HAUSANLP AT SEMEVAL-2023 TASK 12: LEVERAGING AFRICAN LOW RESOURCE TWEETDATA FOR SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present the findings of SemEval-2023 Task 12, a shared task on sentiment analysis for low-resource African languages using twitter dataset. The task featured three subtasks; subtask A is monolingual sentiment classification with 12 tracks which are both monolingual languages, subtask B is multilingual sentiment classification using the tracks in subtask A and subtask C is a zero-shot sentiment classification. We present the result and findings of subtask A and subtask B only. Our goal is to leverage low resource tweet data using Afro-xlmr-large, Bert-base-arabic-camelbert-da-sentiment and Afroxlmr-large-ner-masakhaner-1.0-2.0 models for sentiment analysis of 12 African languages. The dataset for this tasks consists of a gold standard multi-class labelled Twitter dataset from these languages. Our results demonstrates that Afro-xlmr-large model performed better compared to the other models in most of the languages dataset. Similarly, Nigerian languages Hausa, Igbo and Yoruba achieved better performance compared to other languages and this can be attributed to the higher volume of data present in the languages.

1 INTRODUCTION

Social media offers an opinionated platform of data contents on many topics of interest, such as product reviews, feedback on purchases or services, political interest, etc., that is dynamically created by users in different languages. Sentiment analysis in many languages is required due to the need for effective classification of these contents. However, the majority of research on sentiment analysis has been done in high resource languages Jha et al. (2016); Vilares et al. (2016); Heikal et al. (2018); Elhadad et al. (2019); Arora & Kansal (2019); Carvalho & Plastino (2021) while several low-resource African languages receive little attention in Natural Language Processing (NLP) application due to insufficient of publicly available data. Although, recently, some considerable efforts are made for development of sentiment analysis for some low resource African languages Devlin et al. (2018b); Yimam et al. (2020); Ogueji et al. (2021); Abubakar et al. (2021); Muhammad et al. (2022), nonetheless, sentiment analysis for low-resourced African languages still is a misrepresented research area.

The AfriSenti[1] shared task 12 aims at building sentiment analysis for 14 low-resource African languages using Twitter dataset which include Hausa, Yoruba, Igbo, Nigerian Pidgin from Nigeria, Amharic, Tigrinya, and Oromo from Ethiopia, Swahili from Kenya and Tanzania, Algerian Arabic dialect from Algeria, Kinyarwanda from Rwanda, Twi from Ghana, Mozambique Portuguese from Mozambique and Moroccan Arabic/Darija from Morocco. The task featured the following subtasks: Subtask A (Monolingual Sentiment Classification): Given a training document, analyse the sentiment classification of 12 individual African.languages. Subtask B (Multilingual Sentiment Classification): Given a training document, analyse the sentiment classification of multiple languages using the languages in task A. Subtask C (Zero-Shot Sentiment Classification): Given an evaluation data only, analyse the sentiment classification of only two African languages.

In this paper, we demonstrate our approach in tackling the concerns raised in the SemEval Task 12 as well as the obstacles posed by low resource languages. We used Bert-base-multilingual-cased,

[1] AfriSenti-SemEval: Sentiment Analysis for Low-resource African Languages using Twitter Dataset

`AfriBERTa_large` and `Afro-xlmr-large` model to facilitate the classification of tweets using monolingual and multilingual approach.The rest of the paper is organized as follows. Section 2 is the related works. Section 3 describes the proposed approach. Experiment and evaluation are discussed in section 4, while section 5 draws some conclusions and discusses some directions for future work.

2 Related Work

Sentiment analysis (SA) is the process of generating positive or negative sentiment from data source using computational techniques, and to some extend is capable of predicting and classifying temper such as excitement, anger and sadnessFortin & Chaib-Draa (2019). Sentiment and emotion are considered to be related in natural language processing Akhtar et al. (2019). Therefore, the process of assigning polarity to text and emotion among others has become a prevalent task in natural language processing Moore & Barnes (2020). Hence, sentiment analysis is to generate opinion based on a given input provided by users McCann et al. (2018). The current campaign in NLP is that sentiment analysis is extensively adopted for opinion mining to collect information of users from different fields of a particular aspect Zaharia et al. (2020). The information been extracted is not restricted to text but could be image, audio or video. Sentiment analysis objective is to scrutinize a data in order to check and identify if positive or negative sentiment exist through computational techniques, another objective is to predict elusive classes of emoticons like excitement, anger, and sadness Fortin & Chaib-Draa (2019).

Zhang et al. (2020) proposed an Opinion Triplet Extraction (OTE) multi-task learning framework that equally extract aspect terms, opinion terms, and at the same time analyses sentiment dependencies. Bidirectional long-short term memory (Bi-LSTM) was employed jointly with rectified linear unit (ReLU) for multi-tasking where aspects and opinions are extracted with two independent heads and employ a sentiment dependency parser as the third head for the prediction of word-level sentiment dependencies. Bidirectional Encoder Representations from Transformers (BERT) were recently used and showed outstanding performance on multitask learning for classifications Dai et al. (2020); Devlin et al. (2018a). A deep attentive multi-task learning framework to achieve a concrete inter-task and interclasses relationship for sentiment, Humour, Sarcasm, Offensive, and Motivation has been proposed in Chauhan et al. (2020), and the results indicate that the multitask learning improve the performance of the sentiment, Humour, Sarcasm, Offensive, and Motivation for the analysis of memes. It is found that the multitask learning outperform the single task for meme prediction in the multi-modal state. The literature indicates that both philosophically and experimentally, BERT is a strong tool and achieves new, cutting-edge results on different natural language processing tasks, raising the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question-answering Test F1 to 93.2 (1.5 point absolute improvement), and SQuAD v2.0 Test F1 to 83.1. (5.1 point absolute improvement) as presented in Devlin et al. (2018a) .

Although sentiment analysis has been extensively in used in many highly resource languages like English and French just to mention a few, little attention is paid African low-resource languages. Muhammad et al. (2022) presented the first extensive human-annotated Twitter sentiment dataset for the Hausa, Igbo, Nigerian-Pidgin, and Yoruba languages—the four most widely spoken in Nigeria—consisting of roughly 30,000 annotated tweets per language (and 14,000 for Nigerian-Pidgin) and a sizable portion of code-mixed tweets. For these low-resource languages, they suggested text collecting, filtering, processing, and labeling techniques that let us build datasets. On the dataset, they test a variety of pre-trained models and transfer methods. They discovered that the most effective methods are language-specific models and language-adaptive fine-tuning.

3 Proposed Approach

In this section, we describe our proposed approach for the shared task i.e leveraging low resource-tweet data for sentiment analysis of African languages. Our goal is to identify the sentiment classification of low resource African languages using AfriBERTa-large, BERT-base-multilingual-cased and Afro-xlmr-large models.

3.1 MODELS DESCRIPTION

`Afro-xlmr-large`: AfroXLMR-large Alabi et al. (2022) was created by MLM adaptation of XLM-R-large model on 17 African languages (Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Oromo, Nigerian Pidgin, Kinyarwanda, Kirundi, Shona, Somali, Sesotho, Swahili, isiXhosa, Yoruba and isiZulu), covering the major African language families and 3 high resource languages (Arabic, French and English).

`Bert-base-arabic-camelbert-da-sentiment`: Bert-base-arabic-camelbert-da-sentiment Inoue et al. (2021) is a collection of BERT models pre-trained on Arabic texts with different sizes and variants. Arabic Sentiment Tweet Dataset (ASTD) Nabil et al. (2015), an Arabic Speech-Act and Sentiment Corpus of Tweets (ArSAS) Elmadany et al. (2018) and SemEval datasets are used for the fine-tuning the model.

`Afroxlmr-large-ner-masakhaner-1.0-2.0`: masakhane/afroxlmr-large-ner-masakhaner-1.0-2.0 Adelani et al. (2021) is a Named Entity Recognition (NER) model for 21 African languages. Specifically, this model is a Davlan/afro-xlmr-large model that was fine-tuned on an aggregation of African language datasets obtained from two versions of MasakhaNER dataset i.e. MasakhaNER 1.0 and MasakhaNER 2.0. The languages covered are: Amharic, Bambara, Ghomala, Ewe, Fon, Hausa, Igbo, Kinyarwanda, Luganda, Dholuo, Mossi, Chichewa, Nigerian, chShona, Kiswahili, Setswana, Twi, Wolof, isiXhosa, Yorùbá and isiZulu. It has been trained to recognize four types of entities: dates and times (DATE), location (LOC), organization (ORG) and person (PER).

3.2 SHARED TASK DESCRIPTION

The AfriSenti-SemEval Shared Task 12 is based on a collection of Twitter datasets in 14 African languages for sentiment classification. Participants are provided with a training dataset and are required to make a prediction using multiclass sentiment classification. It consists of three sub-tasks: Monolingual sentiment classification, multilingual sentiment classification and zero-shot sentiment classification. In this paper we concentrate on only monolingual and multilingual task with a total of 12 languages.

3.2.1 SUBTASK A: MONOLINGUAL SENTIMENT CLASSIFICATION

For this subtask, we used a single language for sentiment classification. Given training data in a target language, we determine the polarity of a tweet in the target language (positive, negative, or neutral). If a tweet conveys both a positive and negative sentiment, whichever is the stronger sentiment should be chosen. This sub-task consists of 12 African languages: Hausa, Yoruba, Igbo, Nigerian-Pidgin, Amharic, Algerian Arabic, Moroccan Arabic/Darija, Swahili, Kinyarwanda, Twi, Mozambican Portuguese, Xitsonga (Mozambique dialect). For this subtask, the dataset is split into 70% training and 30% validation. We select an individual language, fine-tune the models with the provided training data, tune several hyper-parameters to obtain the optimal performances using only 2 models: Afro-xlmr-large and Bert-base-arabic-camelbert-da-sentiment. Afro-xlmr-large is used on all the languages with the exception Darija and Algerian Arabic. And this is because Afro-xlmr-large model was not trained using the two languages, therefore, we used the Bert-base-arabic-camelbert-da-sentiment.

3.2.2 SUBTASK B: MULTILINGUAL SENTIMENT CLASSIFICATION

Given combined training data from Task-A (Track 1 to 12), we determine the polarity of a tweet in the target language (positive, negative, or neutral). For this subtask, the multilingual dataset is split into three parts: 90% training, 5% each for validation and test set. The three part split allows for more robust evaluation of the model, avoids overfitting and produce a better estimate of the model's generalization performance. We implemented only 1 model: Afroxlmr-large-ner-masakhaner-1.0-2.0. And this is because the model was trained using almost all the African languages.

4 EXPERIMENT AND EVALUATION

This section describes the experimental and evaluation settings for our proposed approach for the SemEval- 2023 Task 12.

4.0.1 DATASET DESCRIPTION

AfriSenti dataset is a collection of Twitter dataset for sentiment analysis of African languages. The dataset used for the AfriSenti-SemEval 2023 Shared task 12 consists of 14 languages: Hausa, Yoruba , Igbo, Nigerian-Pidgin, Amharic, Algerian Arabic, Moroccan Arabic/Darija, Swahili, Kinyarwanda, Twi, Mozambican Portuguese, Setswana, isiZulu, Tigrinya, Xitsonga, and Oromo. The datasets are gold standard with multi-class label (positive, negative and neutral). Each tweet is annotated by three annotators following the annotation guidelines in Mohammad (2016). Table 1 shows the distribution of the languages datasets, sentiment labels and sizes.

Table 1: Dataset Distribution

Task A Languages	**Positive**	**Negative**	**Neutral**	**Total-size**
Hausa(ha)	4687	4573	4912	14172
Igbo(ig)	3084	2600	4508	10192
Naija(pcm)	1808	3241	72	05121
Yorùbá(yo)	3542	1872	3108	08522
Amharic(am)	1332	1548	3104	05984
Algerian Arabic(dz)	417	892	342	01651
Xitsonga(ts)	384	284	136	00804
Kinyarwanda(kr)	899	1146	1257	03302
Mozambican Portuguese(pt)	681	782	1600	03063
Swahili(sw)	1072	547	191	01810
Twi(twi)	1644	1315	522	03481
Darija(ma)	1758	1664	2161	05584
Task B				
Multilingual	20783	20108	22794	63685

4.0.2 EXPERIMENTAL SETTINGS

The 2 subtasks uses different training and testing size. We fine-tune the models with the provided training data, tune several hyper-parameters to obtain the optimal performances as shown in Table Table 2 .

Table 2: Subtasks Parameter Set

Subtask	**Max-Length**	**Batch-Size**	**Epochs**	**Optimizer**	**Learning-Rate**
Subtask A	128	16	5	AdamW	1e-5
Subtask B	150	32	20	AdamW	2e-5

All subtasks will be evaluated using standard evaluation metrics, including precision, recall and weighted F1.

5 RESULTS AND DISCUSSION

Table 3 represent the performance of the 12 African languages evaluated by the metrics: precision, recall and weighted F1. For the monolingual subtask, the best dataset performance is Hausa language with weighted F1 of 81.0% followed by Igbo and Yoruba with 77.0% and 71.9% respectively. The Xitsonga language achieved the least performance with weighted F1 of 50.3%. This higher performance can be attributed to the volume of dataset present in the languages Hausa, Igbo and Yoruba. We further perform ablation study on a different monolingual Hausa dataset as a case study. The first result yielded 76% with 70% of the total dataset sample. However, the performance now increases to 81% after adding the gold label dataset. The Xitsonga language has a smaller volume of data of just 804 training size and thus the lower performance. Afro-xlmr-large model is used in 10 of the languages and achieve a better performance when implemented in languages with larger datasets than in smaller dataset. Therefore, we are of the opinion that there is a correlation between performance and volume of data for a given language.

Table 3: Sentiment Classification Performance on two tasks. Task A - Monolingual and Task B - Multilingual

Monolingual-Task A	**Models**	**Precision**	**Recall**	**W-F1**
Amharic	Afro-xlmr-large	79.5	51.6	57.3
Darija	Bert-base-arabic-camelbert-da-sentiment	60.0	58.0	58.5
Hausa	Afro-xlmr-large	81.0	81.0	81.0
Igbo	Afro-xlmr-large	77.2	77.1	77.0
Yoruba	Afro-xlmr-large	72.6	71.6	71.9
Twi	Afro-xlmr-large	63.4	65.3	64.1
Nigerian Pidgin	Afro-xlmr-large	65.0	72.7	68.5
Algerian Arabic	Bert-base-arabic-camelbert-da-sentiment	65.3	65.9	65.1
Mozambique Portuguese	Afro-xlmr-large	70.6	67.5	68.4
Swahili	Afro-xlmr-large	63.2	63.4	63.2
Kinyarwanda	Afro-xlmr-large	70.6	70.6	70.6
Xitsonga	Afro-xlmr-large	46.9	55.1	50.3
Multilingual-Task B				
Multilingual	Afroxlmr-large-ner-masakhaner-1.0-2.0	69.4	69.4	69.5

And for the multilingual task, Afroxlmr-large-ner-masakhaner-1.0-2.0 model achieved a weighted F1 of 69.5%. This shows that there's a potential for a cross-lingual model for building AI system. To further show that higher volume of data for a language improve the performance of deep learning models, the multilingual result increases to 69.5% from 63.9% after combining the training and development set.

6 CONCLUSION AND FUTUER WORK

In this paper, we presented an overview of the SemEval shared task on sentiment analysis for low-resource African languages using Twitter dataset.We employed Afro-xlmr-large for subtasks A and was implemented by all the language dataset with the exception of Darija and Algerian Arabic that uses Bert-base-arabic-camelbert-da-sentiment. Subtask B uses Afroxlmr-large-ner-masakhaner. Experimental results demonstrate that Afro-xlmr-large model performed better compared to the other models in most of the languages dataset. Similarly, Nigerian languages Hausa, Igbo and Yoruba achieved better performance compared to other languages and this can be attributed to the higher

volume of data present in the languages. Our results indicate that deep learning are effective in sentiment classification in African language given the right data, model and training. For future work, we recommend larger dataset for all the languages for this particular task and active learning can be utilized to speed up annotation process for languages with smaller dataset.

REFERENCES

Amina Imam Abubakar, Abubakar Roko, Aminu Muhammad Bui, and Ibrahim Saidu. An enhanced feature acquisition for sentiment analysis of english and hausa tweets. International Journal of Advanced Computer Science and Applications, 12(9), 2021.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. MasakhaNER: Named Entity Recognition for African Languages. Transactions of the Association for Computational Linguistics, 9:1116–1131, 10 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00416. URL https://doi.org/10.1162/tacl_a_00416.

Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. arXiv preprint arXiv:1905.05812, 2019.

Jesujoba O Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In Proceedings of the 29th International Conference on Computational Linguistics, pp. 4336–4349, 2022.

Monika Arora and Vineet Kansal. Character level embedding with deep convolutional neural network for text normalization of unstructured data for twitter sentiment analysis. Social Network Analysis and Mining, 9:1–14, 2019.

Jonnathan Carvalho and Alexandre Plastino. On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. Artificial Intelligence Review, 54:1887–1936, 2021.

Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. All-in-one: A deep attentive multi-task learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes. In Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing, pp. 281–290, 2020.

Zehui Dai, Cheng Peng, Huajie Chen, and Yadong Ding. A multi-task incremental learning framework with category name embedding for aspect-category sentiment analysis. arXiv preprint arXiv:2010.02784, 2020.

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

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

Mohamed K Elhadad, Kin Fun Li, and Fayez Gebali. Sentiment analysis of arabic and english tweets. In Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33, pp. 334–348. Springer, 2019.

A Elmadany, Hamdy Mubarak, and Walid Magdy. Arsas: An arabic speech-act and sentiment corpus of tweets. OSACT, 3:20, 2018.

Mathieu Pagé Fortin and Brahim Chaib-Draa. Multimodal sentiment analysis: A multitask learning approach. In ICPRAM, pp. 368–376, 2019.

Maha Heikal, Marwan Torki, and Nagwa El-Makky. Sentiment analysis of arabic tweets using deep learning. Procedia Computer Science, 142:114–122, 2018.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. The interplay of variant, size, and task type in arabic pre-trained language models. arXiv preprint arXiv:2103.06678, 2021.

Vandana Jha, GR Shreedevi, P Deepa Shenoy, and KR Venugopal. Generating multilingual subjectivity resources using english language. Int. J. Comput. Appl, 152(9):41–47, 2016.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. arXiv preprint arXiv:1806.08730, 2018.

Saif Mohammad. A practical guide to sentiment annotation: Challenges and solutions. In Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp. 174–179, 2016.

Andrew Moore and Jeremy Barnes. Multi-task learning of negation and speculation for targeted sentiment classification. arXiv preprint arXiv:2010.08318, 2020.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Anuoluwapo Aremu, Saheed Abdul, and Pavel Brazdil. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. arXiv preprint arXiv:2201.08277, 2022.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. Astd: Arabic sentiment tweets dataset. In Proceedings of the 2015 conference on empirical methods in natural language processing, pp. 2515–2519, 2015.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 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, pp. 116–126, 2021.

David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. En-es-cs: An english-spanish code-switching twitter corpus for multilingual sentiment analysis. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 4149–4153, 2016.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models. In Proceedings of the 28th International Conference on Computational Linguistics, pp. 1048–1060, 2020.

George-Eduard Zaharia, George-Alexandru Vlad, Dumitru-Clementin Cercel, Traian Rebedea, and Costin-Gabriel Chiru. Upb at semeval-2020 task 9: Identifying sentiment in code-mixed social media texts using transformers and multi-task learning. arXiv preprint arXiv:2009.02780, 2020.

Chen Zhang, Qiuchi Li, Dawei Song, and Benyou Wang. A multi-task learning framework for opinion triplet extraction. arXiv preprint arXiv:2010.01512, 2020.