UCAS-IIE-NLP at SemEval-2023 Task 12: Enhancing Generalization of Multilingual BERT for Low-resource Sentiment Analysis

Dou Hu[1,2] and **Lingwei Wei**[1,2] and **Yaxin Liu**[1,2] and **Wei Zhou**[1] and **Songlin Hu**[1,2]
[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
{hudou, weilingwei, liuyaxin, zhouwei, husonglin}@iie.ac.cn

Abstract

This paper describes our system designed for SemEval-2023 Task 12: Sentiment analysis for African languages. The challenge faced by this task is the scarcity of labeled data and linguistic resources in low-resource settings. To alleviate these, we propose a generalized multilingual system **SACL-XLMR** for sentiment analysis on low-resource languages. Specifically, we design a lexicon-based multilingual BERT to facilitate language adaptation and sentiment-aware representation learning. Besides, we apply a supervised adversarial contrastive learning technique to learn sentiment-spread structured representations and enhance model generalization. Our system achieved competitive results, largely outperforming baselines on both multilingual and zero-shot sentiment classification subtasks. Notably, the system obtained the **1st** rank on the zero-shot classification subtask in the official ranking. Extensive experiments demonstrate the effectiveness of our system.

1 Introduction

Sentiment analysis is a critical aspect of natural language processing with numerous applications, including public opinion monitoring (Boon-Itt and Skunkan, 2020), healthcare services (Zunic et al., 2020), and recommendation systems (Hu et al., 2021b). However, performing sentiment analysis in low-resource languages poses significant challenges, including the scarcity of labeled data and linguistic resources, as well as the diversity of languages and dialects (Lo et al., 2017; Oueslati et al., 2020). In SemEval-2023 Task 12 (Muhammad et al., 2023b), the focus is on sentiment analysis for African languages in Twitter, which further exacerbates the challenges due to the presence of tone, code-switching, and digraphia phenomena (Adebara and Abdul-Mageed, 2022).

Although multilingual pre-trained language models (multilingual PTMs) (Conneau and Lample, 2019; Conneau et al., 2020) have shown potential in cross-lingual transfer learning compared to monolingual PTMs (Devlin et al., 2019; Hu et al., 2022a), they have limitations in capturing nuances and cultural differences within a language, especially in the context of dialects and regional variations.

In this paper, we propose a generalized multilingual system named **SACL-XLMR** to address these limitations and enhance the generalization of multilingual PTMs for under-represented languages, particularly African languages. Our system leverages a lexicon-based multilingual BERT model to facilitate language adaptation and sentiment-aware representation learning. Additionally, we apply a supervised adversarial contrastive learning (SACL) technique (Hu et al., 2023) to learn sentiment-spread structured representations and enhance model generalization.

We present the details of the proposed system and evaluate its performance on SemEval-2023 Task 12. Our system achieves remarkable performance, outperforming baselines by **+1.1%** weighted-F1 score on multilingual sentiment classification subtask and by **+2.8%** weighted-F1 score on zero-shot sentiment classification subtask in the AfriSenti-SemEval datasets (Muhammad et al., 2023a). Moreover, following the AfriSenti SemEval Prizes[1] and the task description (Muhammad et al., 2023b), our system obtains the **1st** rank on the zero-shot classification subtask in the official ranking. We conducted experiments to demonstrate the effectiveness of our approach, highlighting the potential of our system in overcoming the challenges of low-resource sentiment analysis.

2 Background

2.1 Task and Data Description

The SemEval-2023 Task 12: Sentiment analysis for African languages (AfriSenti-SemEval) (Muham-

[1] https://afrisenti-semeval.github.io/prizes/

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1849–1857
July 13-14, 2023 ©2023 Association for Computational Linguistics

ISO Code	Language	Total	Train	Val	Test	Subregion	Script	Lexicon
amh	Amharic	9,483	5,985	1,498	2,000	East Africa	Ethiopic	✗
arq	Algerian Arabic/Darja	3,062	1,652	415	959	North Africa	Arabic	✗
hau	Hausa	22,155	14,173	2,678	5,304	West Africa	Latin	✓
ibo	Igbo	15,718	10,193	1,842	3,683	West Africa	Latin	✓
kin	Kinyarwanda	5,158	3,303	828	1,027	East Africa	Latin	✓
ary	Moroccan Arabic/Darija	9,762	5,584	1,216	2,962	Northern Africa	Arabic/Latin	✓
pt-MZ	Mozambican Portuguese	7,495	3,064	768	3,663	Southeastern Africa	Latin	✗
pcm	Nigerian Pidgin	10,559	5,122	1,282	4,155	West Africa	Latin	✗
orm	Oromo	2,494	-	397	2,097	East Africa	Latin	✓
swa	Swahili	3,014	1,811	454	749	East Africa	Latin	✗
tir	Tigrinya	2,400	-	399	2,001	East Africa	Ethiopic	✓
twi	Twi	4,821	3,482	389	950	West Africa	Latin	✓
tso	Xitsonga	1,264	805	204	255	Southern Africa	Latin	✗
yor	Yorùbá	15,130	8,523	2,091	4,516	West Africa	Latin	✓

Table 1: The statistics of the AfriSenti datasets. The train/validation sets of Oromo (orm) and Tigrinya (tir) are not used due to the zero-shot transfer setting used for evaluation. Lexicon refers to a valid lexicon, which provides words or phrases that correspond to the predefined sentiment polarity.

mad et al., 2023b) is the first Afro-centric SemEval shared task for sentiment analysis in Twitter. It consists of three subtasks, i.e., monolingual, multilingual, and zero-shot sentiment classification. Brief descriptions of the last two subtasks that our team focuses on are as follows:

- **Multilingual Sentiment Classification**. Given combined training data of multiple African languages, determine the polarity of a tweet on the combined test data of the same languages (positive, negative, or neutral). This subtask has only one track with 12 languages (Amharic, Algerian Arabic/Darja, Hausa, Igbo, Kinyarwanda, Moroccan Arabic/Darija, Mozambican Portuguese, Nigerian Pidgin, Swahili, Twi, Xitsonga, and Yorùbá), i.e., a multilingual track with 12 African languages.

- **Zero-Shot Sentiment Classification**. Given unlabelled tweets in two African languages (Tigrinya and Oromo), leverage any or all available training datasets of source languages (12 African languages in the multilingual track) to determine the sentiment of a tweet in the two target languages. This task has two tracks, i.e., a zero-shot Tigrinya track and a zero-shot Oromo track.

The AfriSenti datasets[2] (Muhammad et al., 2023a) are a collection of multilingual Twitter datasets that consist of 110,000+ tweets in 14 low-resource African languages from four language families for sentiment analysis. The statistics of each monolingual tweet datasets are reported in Table 1. The datasets involve tweets labeled with three sentiment classes (positive, negative, neutral). Each tweet is annotated by three native speakers following the sentiment annotation guidelines Mohammad (2016) and the final label for each tweet is determined by majority voting (Davani et al., 2022). If a tweet conveys both a positive and negative sentiment, the stronger sentiment should be chosen.

3 Related Work

3.1 Sentiment Analysis

Sentiment analysis has evolved from lexicon-based approaches to more advanced machine learning and deep learning-based methods (Medhat et al., 2014). Previous works in sentiment analysis have focused on various levels of granularity, such as aspect (Pontiki et al., 2014), sentence (Hu et al., 2021a), and document (Wei et al., 2020), as well as different modalities (Zadeh et al., 2017; Hu et al., 2022b) and languages (Boiy and Moens, 2009; Balahur and Turchi, 2014).

3.2 Low-resource Sentiment Analysis

Despite the success of polarity classification in high-resource languages, noisy user-generated data in under-represented languages presents a challenge (Yimam et al., 2020). Recently, several studies have proposed approaches for sentiment analysis on low-resource languages (Lo et al., 2017; Yimam et al., 2020). Besides, Moudjari et al. (2020); Adebara and Abdul-Mageed (2022); Muhammad

[2] https://github.com/afrisenti-semeval

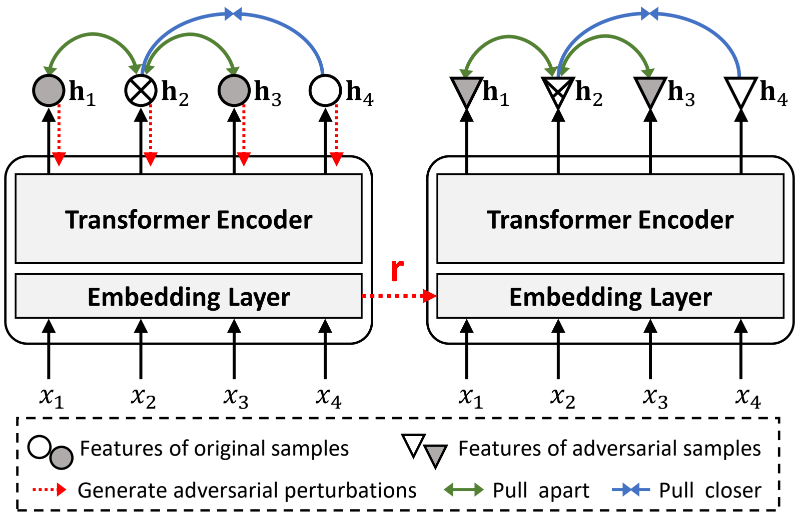

Figure 1: Overall architecture of our SACL-XLMR. Given a batch of training samples, a multilingual BERT is used to learn contextual representations of the input sentences. We take the $\times$-marked utterance as an example to show the objective of SACL. r means adversarial perturbations that put on the embedding layer of BERT.

et al. (2023a) have relied on manual annotation by native speakers or expert annotators to build sentiment analysis datasets in low-resource languages.

4 System Overview

In this section, we describe our system adopted in SemEval-2023 Task 12, where we design a generalized multilingual system named **SACL-XLMR** for sentiment analysis on low-resource languages. The overall architecture is illustrated in Figure 1. With the guidance of SACL framework, the method can learn label-consistent structured features for better sentiment classification.

4.1 Model Architecture

The network structure of SACL-XLMR consists of a multilingual BERT (i.e., an embedding layer and Transformer encoder) and a sentiment classifier.

Multilingual BERT We apply a multilingual BERT model (Conneau and Lample, 2019; Alabi et al., 2022) on monolingual corpus to facilitate language adaptation. Besides, sentiment lexicon knowledge for each language is used to enhance sentiment-aware representation learning.

Formally, given an input token sequence $x_{i1}, ..., x_{iN}$ where x_{ij} refers to j-th token in the i-th input sample, and N is the maximum sequence length, the model learns to generate the context representation of the input token sequences:

$$\mathbf{h}_i = \mathrm{BERT}(\texttt{[CLS]}, s_L, \texttt{[SEP]}, x_{i1}, ..., x_{iN}, \texttt{[SEP]}), \quad (1)$$

where `[CLS]` and `[SEP]` are special tokens, usually at the beginning and end of each sequence, respectively. s_L refers to a token sequence of sentiment lexicon prefix corresponding to the input sequence. $\mathbf{h}_i$ indicates the hidden representation of the i-th input sample, computed by the representation of `[CLS]` token in the last layer of the encoder.

Sentiment Classifier Finally, according to the obtained representations, a sentiment classifier is applied to predict the sentiment label of each sample.

$$\hat{\mathbf{y}}_i = softmax(\mathbf{W}_h \mathbf{h}_i + \mathbf{b}_h), \quad (2)$$

where $\mathbf{W}_h \in \mathbb{R}^{d_h \times |\mathcal{Y}|}$ and $\mathbf{b}_h \in \mathbb{R}^{|\mathcal{Y}|}$ are trainable parameters. $|\mathcal{Y}|$ is the number of sentiment labels.

4.2 Optimization Objective

Supervised contrastive learning (SCL) (Khosla et al., 2020; Gunel et al., 2021) is utilized to learn a generalized feature representation by capturing similarities between examples within a class and contrasting them with examples from other classes. However, directly compressing the feature space of each class can harm fine-grained features, which limits the model's ability to generalize. Recently, a supervised adversarial contrastive learning (SACL) technique (Hu et al., 2023) is proposed to address this issue by learning class-spread structured representations in a supervised manner. It can effectively utilize prior information on label consistency and retain fine-grained intra-class features.

Model	# Param.	# Vocab.	# Lang.	Seen Lang. Adapt.	Unseen Lang. Adapt.	Lang. supported in AfriSenti datasets
XLM-R	270M	250k	100	✗	✗	amh, arq, hau, ary, pt-MZ, orm, swa
AfriBERTa	126M	70k	11	✓	✗	amh, hau, ibo, kin, pcm, orm, swa, tir, yor
AfroXLMR	270M	250k	20	✓	✗	amh, arq, hau, ibo, kin, ary, pcm, orm, swa, yor
SACL-XLMR	270M	250k	20	✓	✓	amh, arq, hau, ibo, kin, ary, pcm, orm, swa, yor

Table 2: Comparison of our SACL-XLMR with other PTMs. # Param. refers to the total number of parameters for each model excluding the task-specific classifier. # Vocab. represents the size of vocabulary. # Lang. indicates the number of language coverage. Seen/Unseen Lang. Adapt. represents whether the model supports seen/unseen target language adaptation. We list the languages covered by both the pre-trained corpus and AfriSent datasets.

In this task, we apply the SACL technique to learn sentiment-spread representations and enhance the generalization of multilingual BERT. Formally, let us denote I as the set of samples in a batch. Define $\phi(i) = \{e \in I \backslash \{i\} : \hat{\mathbf{y}}_e = \hat{\mathbf{y}}_i\}$ as the set of indices of all positives in the batch distinct from i, and $|\phi(i)|$ is its cardinality. The loss function of soft SCL is a weighted average of CE loss and SCL loss with a trade-off scalar parameter λ, i.e.,

$$\mathcal{L}_{\text{soft-SCL}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{SCL}}, \tag{3}$$

where

$$\mathcal{L}_{\text{CE}} = -\sum_{i \in I} \mathbf{y}_{i,k} \log(\hat{\mathbf{y}}_{i,k}), \tag{4}$$

$$\mathcal{L}_{\text{SCL}} = \sum_{i \in I} \frac{-1}{|\phi(i)|} \sum_{e \in \phi(i)} \log \frac{\exp(sim(\mathbf{z}_i, \mathbf{z}_e)/\tau)}{\sum\limits_{a \in A(i)} \exp(sim(\mathbf{z}_i, \mathbf{z}_a)/\tau)}. \tag{5}$$

$\mathbf{y}_{i,k}$ and $\hat{\mathbf{y}}_{i,k}$ denote the value of one-hot vector $\mathbf{y}_i$ and probability vector $\hat{\mathbf{y}}_i$ at class index k, respectively. $A(i) = I \backslash \{i\}$. $\mathbf{z}_i = \mathbf{W}_h \mathbf{h}_i + \mathbf{b}_h$. $sim(\cdot, \cdot)$ is a pairwise similarity function, i.e., dot product. $\tau > 0$ is a scalar temperature parameter that controls the separation of classes.

At each step of training, under the soft SCL objective, we apply an adversarial training strategy (e.g., FGM (Miyato et al., 2017)) on original samples to generate adversarial samples. These samples can be seen as hard positive examples, which spread out the representation space for each sentiment class and confuse robust-less models. After that, we utilize a new soft SCL on obtained adversarial samples to maximize the consistency of sentiment-spread representations with the same sentiment label. Following the above calculation process of $\mathcal{L}_{\text{soft-SCL}}$ on original samples, the optimization objective on corresponding adversarial samples can be easily obtained in a similar way, i.e., $\mathcal{L}^{\text{r-adv}}_{\text{soft-SCL}}$.

The overall loss of SACL is defined as a sum of two soft SCL losses on both original and adversarial samples, i.e.,

$$\mathcal{L} = \mathcal{L}_{\text{soft-SCL}} + \mathcal{L}^{\text{r-adv}}_{\text{soft-SCL}}. \tag{6}$$

5 Experimental Setup

5.1 Comparison Methods

We compare SACL-XLMR with the following several methods:

- **Random** is based on random guessing, choosing each class/label with an equal probability.
- **XLM-R** (Conneau and Lample, 2019) is a multilingual variant of RoBERTa (Liu et al., 2019). It is pre-trained on filtered CommonCrawl data containing 100 languages. We use *xlm-roberta-base*[3] to initialize XLM-R.
- **AfriBERTa** (Ogueji et al., 2021) is an Afrocentri multilingual language model pretrained on 11 African languages. It is trained on an aggregation of datasets from the BBC news website and Common Crawl. We use *castorini/afriberta_large*[3] to initialize AfriBERTa.
- **AfroXLMR** (Alabi et al., 2022) is an XLM-R model adapted to African languages. It is obtained by MLM adaptation of XLM-R on 17 African languages covering the major African language families and 3 high resource languages (Arabic, French, and English). We use *Davlan/afro-xlmr-large*[3] to initialize AfroXLMR.

We report the comparison of our SACL-XLMR and the above PTMs in Table 2.

5.2 Implementation Details

All experiments are conducted on a single NVIDIA Tesla V100 32GB card. Stratified k-fold cross validation (Kohavi, 1995) is performed to split combined training and validation data of 12 African

[3]https://huggingface.co/

Hyperparameter	SACL-XLMR
Hidden size d_u	1024
Perturbation radius	$\{0.5, 5\}$
Perturbation rate	$\{0.1, 1\}$
Trade-off weight λ and $\lambda^{\text{r-adv}}$	$\{0.05, 0.1\}$
Temperature τ and $\tau^{\text{r-adv}}$	0.1
Number of epochs	10
Patience	3
Batch size	128
Learning rate	$1e^{-5}$
Weight decay	$1e^{-2}$
Dropout	0.2
Maximum token length	250

Table 3: Hyperparameter settings of SACL-XLMR.

languages into 5 folds. Train/validation sets for Oromo (`orm`) and Tigrinya (`tir`) are not used due to the limited size of the data. We only evaluate on them in a zero-shot transfer setting. We choose the optimal hyperparameter values based on the the average result of validation sets for all folds, and evaluate the performance of our system on the test data. Following the scoring program of AfriSenti-SemEval, we report the weighted-F1 (w-F1) score to measure the overall performance.

Our SACL-XLMR is initialized with the *Davlan/afro-xlmr-large*[3] parameters, due to the nontrivial and consistent performance in both subtasks. The network parameters are optimized by using Adam optimizer (Kingma and Ba, 2015). The class weights in CE loss are applied to alleviate the class imbalance problem and are set by their relative ratios in the train and validation sets. The detailed experimental settings on both two subtasks are in Table 3.

To effectively utilize sentiment lexicons of partial languages in the AfriSenti datasets, we concatenate the corresponding lexicon prefix with the original input text. Given the i-th input sample, the lexicon prefix can be represented as $y_k : w_{k1}, ..., w_{kM}$ where y_k is the sentiment label, w_{km} refers to the corresponding m-th lexicon token in the original sequence. For our final system, we only use sentiment lexicons on the zero-shot subtask. We do not use it on the multilingual subtask due to the fact that some languages in the multilingual target corpus do not have available sentiment lexicons, making it difficult for the model to adapt effectively.

Model	`multilingual`
Random	33.3
XLM-R	62.5
AfriBERTa	64.5
AfroXLMR	69.9
SACL-XLMR$_{fold1}$†	70.3
SACL-XLMR	**71.0**
Improve	+1.1%

Table 4: Experimental results (%) against various methods on the multilingual sentiment classification subtask. We present the weighted-F1 score to measure the performance. All compared pre-trained models are fine-tuned on the multilingual dataset. $fold1$ means the result using only training data of one fold. † indicates the results on the official ranking.

Model	`tir`	`orm`	Avg.
Random	34.3	33.6	34.0
XLM-R	43.8	35.9	39.9
AfriBERTa	44.1	43.6	43.9
AfroXLMR	69.8	42.3	56.1
SACL-XLMR$_{fold1}$†	70.5	45.8	58.2
SACL-XLMR	**71.8**	**46.0**	**58.9**
Improve	+2.0%	+2.4%	+2.8%

Table 5: Experimental results (%) against various methods on the zero-shot sentiment classification subtask. We present the weighted-F1 score to measure the performance. All compared pre-trained models are fine-tuned on the multilingual dataset. $fold1$ means the result using only training data of one fold. † indicates the results on the official ranking.

6 Results and Analysis

6.1 Overall Results

The overall results for both subtasks are summarized in Table 4 and 5. From the results, it is not surprising that all pre-trained models clearly outperformed the Random baseline. The proposed SACL-XLMR consistently outperformed the comparison methods on both subtasks. Specifically, SACL-XLMR achieved **1.1%** and **2.8%** absolute improvements on the multilingual and zero-shot sentiment classification subtasks, respectively.

Moreover, we present the official results from several top-ranked systems for the zero-shot sentiment classification subtask in AfriSenti-SemEval Shared Task (i.e., SemEval-2023 Task 12) in Table 6. Our submitted system obtained the **1st** overall rank on the zero-shot sentiment classification subtask in the official ranking.

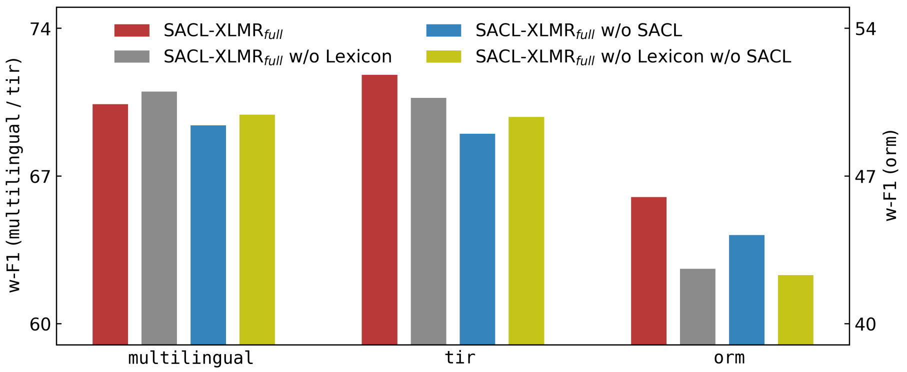

Figure 2: Ablation study results on two subtasks. We report the weighted-F1 score.

Overall Rank	Team Name	tir	orm	Avg.
Top 1	**UCAS-IIE-NLP**	70.47	45.82	**58.15**
Top 2	NLNDE	**70.86**	44.97	57.92
Top 3	ymf924	70.39	45.34	57.87
-	UM6P	69.53	45.27	57.40
-	TBS	69.61	45.12	57.37
-	uid	69.90	44.75	57.33
-	mitchelldehaven	66.96	**46.23**	56.60

Table 6: Results of our submitted system compared with several top-ranked systems for the zero-shot sentiment classification subtask in AfriSenti-SemEval Shared Task. The official scoring program uses the weighted-F1 score to measure the performance. Following the AfriSenti SemEval Prizes[1] and the task description (Muhammad et al., 2023b), the overall rank is calculated by averaging the results of all the languages in the subtask.

6.2 Ablation Study

In this part, we conduct ablation studies by removing key components of SACL-XLMR$_{full}$ to further understand the proposed model:

- **- w/o Lexicon** refers to removing the sentiment lexicon.
- **- w/o SACL** means replacing the SACL objective with a simple cross-entropy (CE) term.
- **- w/o Lexicon - w/o SACL** indicates removing sentiment lexicons and replacing the SACL objective with a CE term, degenerated to AfroXLMR.

Figure 2 shows results of ablation studies on two subtasks for low-resource sentiment analysis. Our SACL-XLMR$_{full}$ w/o Lexicon and SACL-XLMR$_{full}$ yield the best performance on multilingual and zero-shot sentiment classification subtasks, respectively. When removing the SACL objective and replacing it with a CE term, the results consistently decline on all subtasks, showing the effectiveness of SACL.

For the multilingual sentiment classification subtask, SACL-XLMR$_{full}$ obtains sub-optimal results. This is most likely due to the fact that some languages in the target corpus do not have available sentiment lexicons, making it difficult for the model to adapt effectively. Also, another caused factor is the incompleteness and poor quality of lexicon. For the zero-shot sentiment classification subtask, the SACL-XLMR$_{full}$ yields the best performance on both tir and orm languages. It shows the effectiveness of sentiment lexicons in zero-shot scenarios, even if its quality is not good enough.

6.3 Error Analysis

Figure 3 shows an error analysis of our system on two subtasks of AfriSenti-SemEval, including a multilingual test set and two zero-shot test sets. The normalized confusion matrices are used to evaluate the quality of the predicted outputs of SACL-XLMR.

From the diagonal elements of the matrices, true positives of non-neutral labels exceed those of the neutral label. The results show that positive and negative features are more likely to adapt to low-resource languages. Besides, the above phenomenon is more obvious for tir and orm languages. It indicates that SACL-XLMR can further facilitate language adaptation for low-resource languages by making full use of existing sentiment lexicons which contain only positive and negative words.

The confusion matrix of SACL-XLMR reveals the most confusing pair of sentiment labels: neutral to negative, especially for tir and orm languages in a zero-shot setting. The performance on orm language is relatively poor. Apart from the com-

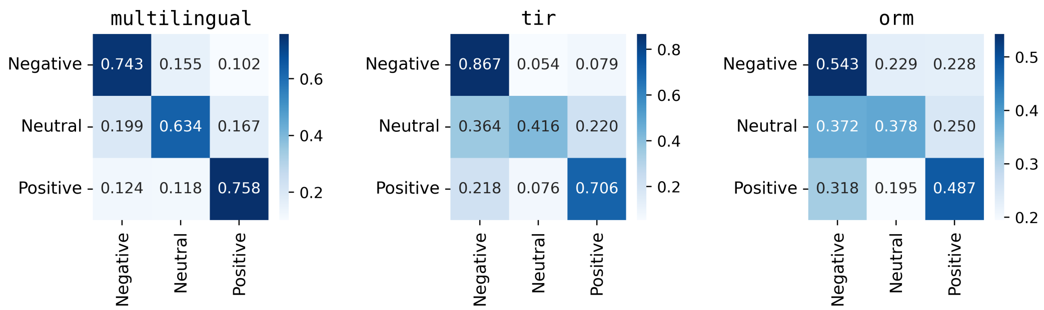

Figure 3: The normalized confusion matrices for SACL-XLMR on three test sets of AfriSenti. The rows represent the actual sentiment labels, whereas the columns represent predictions made by the model. Each cell (i, j) represents that the percentage of class i was predicted as class j. The values of the diagonal elements represent the degree of correctly predicted classes. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

plexity of the language and the phenomenon of data scarcity, it is also due to the significant differences between `orm` and other African languages. Considering the above issues make the task optimization more difficult, there is still a lot of room for improvement.

7 Conclusion

In this paper, a multilingual system named SACL-XLMR has been proposed for sentiment analysis on low-resource African languages. The system employs a lexicon-based multilingual BERT to facilitate language adaptation and sentiment-aware representation learning. It also uses a supervised adversarial contrastive learning technique to learn sentiment-spread structured representations and enhance model generalization. The system achieved competitive results, largely outperforming the comparison baselines on both multilingual and zero-shot sentiment classification subtasks, and obtained the 1st rank on zero-shot classification subtask in the official ranking.

Acknowledgements

All the work in this paper are conducted during the SemEval-2023 Competition. We thank the SemEval-2023 organizers and AfriSenti-SemEval task organizers for making this research possible. We also appreciate the anonymous reviewers for their insightful and constructive comments that have helped us improve the quality of the paper.

References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 4336–4349. International Committee on Computational Linguistics.

Alexandra Balahur and Marco Turchi. 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Comput. Speech Lang.*, 28(1):56–75.

Erik Boiy and Marie-Francine Moens. 2009. A machine learning approach to sentiment analysis in multilingual web texts. *Inf. Retr.*, 12(5):526–558.

Sakun Boon-Itt and Yukolpat Skunkan. 2020. Public perception of the covid-19 pandemic on twitter: Sentiment analysis and topic modeling study. *JMIR Public Health Surveill*, 6(4):e21978.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7057–7067.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Trans. Assoc. Comput. Linguistics*, 10:92–110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *The 61st Annual Meeting of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Dou Hu, Xiaolong Hou, Xiyang Du, Mengyuan Zhou, Lianxin Jiang, Yang Mo, and Xiaofeng Shi. 2022a. VarMAE: pre-training of variational masked autoencoder for domain-adaptive language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6276–6286. Association for Computational Linguistics.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lian-Xin Jiang, and Yang Mo. 2022b. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 7037–7041. IEEE.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021a. DialogueCRN: contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7042–7052. Association for Computational Linguistics.

Dou Hu, Lingwei Wei, Wei Zhou, Xiaoyong Huai, Zhiqi Fang, and Songlin Hu. 2021b. PEN4Rec: preference evolution networks for session-based recommendation. In *Knowledge Science, Engineering and Management - 14th International Conference, KSEM 2021, Tokyo, Japan, August 14-16, 2021, Proceedings, Part I*, volume 12815 of *Lecture Notes in Computer Science*, pages 504–516. Springer.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, pages 1137–1145. Morgan Kaufmann.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Siaw Ling Lo, Erik Cambria, Raymond Chiong, and David Cornforth. 2017. Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artif. Intell. Rev.*, 48(4):499–527.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Saif Mohammad. 2016. 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*, pages 174–179, San Diego, California. Association for Computational Linguistics.

Leila Moudjari, Karima Akli-Astouati, and Farah Benamara. 2020. An Algerian corpus and an annotation platform for opinion and emotion analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1202–1210, Marseille, France. European Language Resources Association.

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.

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.

Oumaima Oueslati, Erik Cambria, Moez Ben HajHmida, and Habib Ounelli. 2020. A review of sentiment analysis research in arabic language. *Future Gener. Comput. Syst.*, 112:408–430.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

Lingwei Wei, Dou Hu, Wei Zhou, Xuehai Tang, Xiaodan Zhang, Xin Wang, Jizhong Han, and Songlin Hu. 2020. Hierarchical interaction networks with rethinking mechanism for document-level sentiment analysis. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, Proceedings, Part III*, volume 12459 of *Lecture Notes in Computer Science*, pages 633–649. Springer.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ali Ayele, and Chris Biemann. 2020. Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1048–1060. International Committee on Computational Linguistics.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1103–1114. Association for Computational Linguistics.

Anastazia Zunic, Padraig Corcoran, and Irena Spasic. 2020. Sentiment analysis in health and well-being: Systematic review. *JMIR Med Inform*, 8(1):e16023.