TechSSN at SemEval-2023 Task 12: Monolingual Sentiment Classification in Hausa Tweets

Nishaanth Ramanathan, Rajalakshmi Sivanaiah, Angel Deborah Suseelan and Mirnalinee Thanka Nadar Thanagathai Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, Tamil Nadu, India nishaanth2010302@ssn.edu.in, rajalakshmis@ssn.edu.in, angeldeborahs@ssn.edu.in, mirnalineett@ssn.edu.in

Abstract

This paper elaborates on our work in designing a system for SemEval 2023 Task 12: AfriSenti-SemEval, which involves sentiment analysis for low-resource African languages using the Twitter dataset. We utilised a pre-trained model to perform sentiment classification in Hausalanguage tweets. We used a multilingual version of the roBERTa model, which is pretrained on 100 languages, to classify sentiments in Hausa. To tokenize the text, we used the AfriBERTa model, which is specifically pretrained on African languages.

1 Introduction

Sentiment analysis is a subset of Natural Language Processing (NLP), in which the intent is to classify the sentiment present in textual data (Kincl et al., 2016). It is mostly done in languages that have high resources but not in low-resource languages like Hausa, Yoruba, Igbo, and other African languages, despite the fact that they are being used by nearly one-third of human beings. It is essential for us to respect every language and perform NLP research in all languages. Performing NLP research in all languages is crucial for both political and business reasons (Polpinij, 2014). The rapid growth of the Internet and usage of social platforms like Facebook, Twitter, etc. have expanded the area of NLP research (Martin et al., The task we participated in involved 2021). performing sentiment classification in tweets of an African language, and we chose to focus on the Hausa language (Muhammad et al., 2023b). In this task, tweets must be classified as positive, negative, or neutral. Performing sentiment analysis in low-resource languages like Hausa is crucial not only because it is used by many people but also because of the need to interpret the sentiment of every single person.

Our strategy involved tokenizing textual data, fine-tuning a pre-trained model with the tokenized data, and measuring the model's performance on the development dataset. We fine-tuned a model (Muhammad et al., 2022) that was pre-trained on the NaijaSenti corpus (Muhammad et al., 2022), which was provided in the task overview. Our strategy involved tokenizing textual data,

our strategy involved tokenizing textual data, fine-tuning a pre-trained model with the tokenized data, and measuring the model's performance on the development dataset.

Our system ranked 13^{th} among the 35 submissions in the track we participated in, with a weighted F1 score of 80.32%. The top-ranked team achieved a score of 82.62%. Although our model performed excellently on training and development data, it did not perform well on the test datasets, i.e., it could not make accurate predictions on new data.

2 Background

Task A: monolingual sentiment classification was the sub-task in which we participated. The task required us to determine the polarity (positive, negative, or neutral) of tweets in a target low-resource language, Hausa, through sentiment classification. The dataset(Muhammad et al., 2023a) comprised IDs, tweet texts in Hausa, and their corresponding labels. The training and development datasets were of sizes 14,172 and 2,677, respectively. The test dataset consisted of 5,303 samples. Figure 1 shows examples of the dataset.

Table 1 shows the distribution of the data in the training and development dataset across the various labels (positive, negative and neutral). As the table shows, the dataset comprised of more or less equal number of samples in each class label and thus data augmentation was not necessary.

ID	text	label
ha_train_00001	@user Da kudin da Arewa babu wani abin azo aga	negative
ha_train_00002	@user Kaga wani Adu ar Banda 💔 🔞 wai a haka Shi	negative
ha_train_00003	@user Sai haquri fa yan madrid daman kunce cha	negative
ha_train_00004	@user Hmmm yanzu kai kasan girman allah daxaka	negative
ha_train_00005	@user @user Wai gwamno nin Nigeria suna afa kw	negative

Figure 1: Examples of data set

Label	Train	Dev	Test
Positive	4687	887	1755
Negative	4573	894	1759
Neutral	4912	896	1789
Total	14172	2677	5303

Table 1: Data distribution

3 Related Work

Sentiment analysis or classification is a topic that has been researched for many years for highresource languages like English. Recently it gained attention for the low-resource languages also. A survey of the neural techniques used for lowresource languages are discussed by (Ranathunga et al., 2023). Models are trained for low-resource languages like Persian, Urdu, Indonesian, Arabic, Tamil, Marathi, Hindi, Malayalam etc., by many researchers. Deep learning models, transformers models or transfer learning approaches are mainly used for most of the situations. The comparison of lexicon based approach and bert based approaches for Italian language is done by (Catelli et al., 2022) and the advantages of Bert based models are discussed.

Long Short Term Memory model is used to analyze four thousand Indonesian tweets and achieved an accuracy of 73.2% in (Le et al., 2016). The issues in collecting large corpus for low resource languagues are addressed in (Ekbal and Bhattacharyya, 2022). They have used a deep multi-task multilingual adversarial framework to analyze the sentiments in Hindi with 60% accuracy in movie reviews and 72.14% in product reviews dataset. Multi-Task Text Graph Convolutional Networks is used for sentiment classification in Telugu language by (Marreddy et al., 2022). Recently we worked on structured sentiment analysis in (Anantharaman et al., 2022) and emotion analysis for Tamil language in (S et al., 2022). Multilingual pretrained language model with adaptive fine-tuning is used for African languages by (Alabi et al., 2022).

4 System Overview

The first step we took was the tokenization of the training dataset. This involved converting textual data into numerical data since textual data cannot be fed into the model. We used the Transformers library to perform both tokenization and fine-tuning. Tokenization was carried out using a pre-trained tokenizer obtained from a model called "naija-twitter-sentiment-afriberta-large" (Muhammad et al., 2022) which is available on Hugging Face. We then fine-tuned the XLM-RoBERTa-Large model, which had been pre-trained on the NaijaSenti corpus (Muhammad et al., 2022). Finally, we made predictions using the development dataset and measured the model's performance. Figure 2 shows the distribution of various tweet labels in 14 different African languages. Since the Hausa language (hau) had the highest number of samples, we used that dataset for sentiment classification.

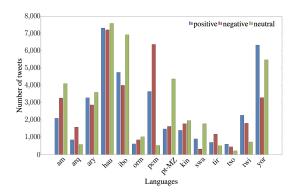


Figure 2: Data distribution of African languages (Muhammad et al., 2023a)

5 Experimental Setup

We trained the model on the entire training dataset, as separate development and testing datasets were available. Prior to training, we mapped the labels in the dataset ('positive', 'negative', and 'neutral') to integers. The dataset was tokenized and dynamically padded. We used a pretrained XLM-RoBERTa-large model (Conneau et al., 2019) to fine-tune the Hausa dataset. XLM-RoBERTa is a multilingual version of the RoBERTa transformer model, pre-trained on 100 different languages with masked language modeling (MLM). Figure 3 illustrates the architecture of the XLM-RoBERTa model (Ranasinghe and Zampieri, 2020).

The XLM-RoBERTa-large model consists of around 125M parameters, 12 layers, 768 hidden

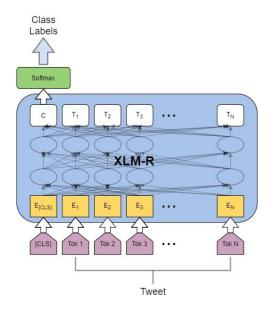


Figure 3: XLM-RoBERTa model (Ranasinghe and Zampieri, 2020)

states, 3072 feed-forward hidden states, and 8 heads. It takes an input sequence of no more than 512 tokens and outputs the sequence representation. The first token of the sequence is always [CLS], which contains the special classification embedding. We used the Adam optimizer and a linear scheduler with an initial learning rate of 2e-5. The model was trained for 10 epochs with a batch size of 32. Given the large number of parameters in XLM-RoBERTa-large and the model's strong performance despite a high batch size, the batch size was set to 32. We used the weighted F1 score as the evaluation metric. The following is a list of the required libraries.

- Transformers ¹
- Pandas²
- NumPy ³
- Scikit-learn⁴

6 Results

Our system achieved weighted F1 scores of 99.94% and 80.07% on the training and development gold data sets, respectively. In the competition, our system ranked 13^{th} with a weighted F1 score of 80.32% on the test data set. Figure 4 shows the confusion matrix for the development data set.

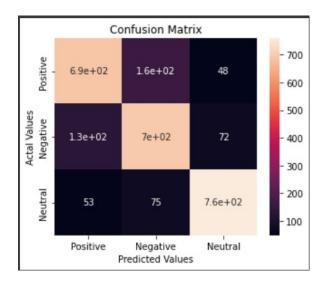


Figure 4: Confusion matrix for development data set

Sentiments	Precision	Recall
Positive	0.79	0.77
Negative	0.75	0.78
Neutral	0.86	0.86

Table 2: Precision and Recall for all three labels

Table 2 shows precision and recall for all three labels for the development dataset.

7 Conclusion

We have fine-tuned a model for Hausa that was pre-trained on four African languages. On the test dataset, we achieved an F1 score of 80.32%, while the top-ranked team achieved 82.62%. Eventhough the model performed better in training, for the new, unseen data, it has not performed well. Therefore, it is essential to provide the model with more training data to prevent overfitting and improve its ability to predict test data. The score can be increased by fine-tuning the hyperparameters and increase the number of epochs. We intend to work on improving the pre-processing stage to get better results and also expand our work into other languages. Pre-processing steps, such as stemming and lemmatization, can be used to enhance performance. Additionally, negating and intensifying words should be handled with care as they can affect the text's sentiment. An approach based on lexicons, which involves assigning sentiment scores to words based on the collection of sentiment lexicons and their corresponding sentiments, can be considered in the pre-processing pipeline. This approach may yield better results.

¹Version: 4.26.1 https://huggingface.co/transformers

²Version: 1.3.5 https://pandas.pydata.org/

³Version: 1.21.6 https://numpy.org/

⁴Version: 1.0.2 https://scikit-learn.org/stable/

References

- Marvin Matías Agüero Torales et al. 2022. Machine learning approaches for topic and sentiment analysis in multilingual opinions and low-resource languages: From english to guarani.
- Jesujoba O Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pretrained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349.
- Karun Anantharaman, Divyasri K, Jayannthan Pt, Angel S, Rajalakshmi Sivanaiah, Sakaya Milton Rajendram, and Mirnalinee T T. 2022. SSN_MLRG1 at SemEval-2022 task 10: Structured sentiment analysis using 2-layer BiLSTM. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1324–1328, Seattle, United States. Association for Computational Linguistics.
- Rosario Catelli, Serena Pelosi, and Massimo Esposito. 2022. Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian. *Electronics*, 11(3):374.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.
- Asif Ekbal and Pushpak Bhattacharyya. 2022. Exploring multi-lingual, multi-task, and adversarial learning for low-resource sentiment analysis. *Transactions on Asian and Low-Resource Language Information Processing*, 21(5):1–19.
- Tomáš Kincl, Michal Novák, and Jiri Pribil. 2016. Sentiment Classification in Multiple Languages: Fifty Shades of Customer Opinions, pages 267–275.
- Tuan Anh Le, David Moeljadi, Yasuhide Miura, and Tomoko Ohkuma. 2016. Sentiment analysis for low resource languages: A study on informal Indonesian tweets. In *Proceedings of the 12th Workshop* on Asian Language Resources (ALR12), pages 123– 131, Osaka, Japan. The COLING 2016 Organizing Committee.
- Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022. Multi-task text classification using graph convolutional networks for large-scale low resource language. In 2022 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE.
- Gati L. Martin, Medard E. Mswahili, and Young-Seob Jeong. 2021. Sentiment classification in swahili language using multilingual bert.

- 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.
- Jantima Polpinij. 2014. Multilingual sentiment classification on large textual data. In 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, pages 183–188.
- Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with crosslingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844, Online. Association for Computational Linguistics.
- Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.
- Varsini S, Kirthanna Rajan, Angel S, Rajalakshmi Sivanaiah. Sakaya Milton Rajendram. and Mirnalinee T. 2022. T Varsini and Kirthanna@DravidianLangTech-ACL2022-emotional analysis in Tamil. In Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 165-169, Dublin, Ireland. Association for Computational Linguistics.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew 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*, pages 1048– 1060, Barcelona, Spain (Online). International Committee on Computational Linguistics.