Witcherses at SemEval-2023 Task 12: Ensemble Learning for African Sentiment Analysis

Monil Gokani **K V Aditya Srivatsa** **Radhika Mamidi**
Language Technologies Research Center (LTRC)
Kohli Center on Intelligent Systems
International Institute of Information Technology, Hyderabad
{monil.gokani, k.v.aditya}@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract

This paper describes our system submission for SemEval-2023 Task 12 AfriSenti-SemEval: Sentiment Analysis for African Languages. We propose an XGBoost-based ensemble model trained on emoticon frequency-based features and the predictions of several statistical models such as SVMs, Logistic Regression, Random Forests, and BERT-based pre-trained language models such as AfriBERTa and AfroXLMR. We also report results from additional experiments not in the system. Our system achieves a mixed bag of results, achieving a best rank of 7th in three of the languages - Igbo, Twi, and Yoruba.

1 Introduction

The AfriSenti Shared Task (Muhammad et al., 2023b) is aimed at promoting sentiment classification research in a diverse group of African Languages. Though sentiment classification is an extremely popular task in NLP, there has been relatively little work done for it in African Languages. Apart from the NaijaSenti dataset (Muhammad et al., 2022) (which is a part of the task itself), some of the languages where such work has been done include Amharic (Yimam et al., 2020), Tunisian Arabizi (Fourati et al., 2021), and Swahili (Martin et al., 2021). The task itself is divided into three sub-tasks - monolingual classification in 12 languages separately, multilingual classification, and zero-shot classification for two languages.

We participated in the monolingual and multilingual tracks of the task. Our system consists of an ensemble model that leverages emoticon frequencies and the predictions of contextualised, transformer-based models such as AfriBERTa (Ogueji et al., 2021) and AfroXLMR (Alabi et al., 2022), as well as simpler models such as Logistic Regression, SVM and Random Forest classifiers. We also release the code on GitHub.[1]

[1] https://github.com/MG1800/afrisenti-ensemble

Our model achieves a significant range of ranks, with our best result being ranked 7th each in the Igbo, Twi, and Yoruba tracks and between 12 and 15 on the multilingual, Hausa and Xitsonga tracks. The full rankings are reported in Table 2.

We report observations on the performances of our individual models that we used across languages and the performance of languages across models. Additionally, we report the results for some of the experiments we carried out that did not make it to the final model.

2 Background

2.1 Task Description

The objective of the task (Muhammad et al., 2023b) is to identify the polarity of a tweet (negative, positive, or neutral) in a set of 14 African languages. It is divided into three sub-tasks.

Sub-Task A is for monolingual classification systems, where a separate classifier is trained for each language and has a separate track for each of the 12 languages in this sub-task. These languages are Hausa, Yoruba, Igbo, Nigerian Pidgin, Amharic, Algerian Arabic, Moroccan Arabic(Darija), Swahili, Kinyarwanda, Twi, Mozambican Portuguese, and Xitsonga(Mozambique Dialect). Sub Task B consists of training a single multilingual model to classify tweets in all of the languages in Sub Task A. Sub Task C is for zero-shot classification in 2 languages (Tigrinya and Oromo), for which training data is not available, and only a development set is provided.

We are participating in all the tracks for Sub Task A and in Sub Task B.

2.2 Dataset

The AfriSenti dataset (Muhammad et al., 2023a), provided by the organisers, consists of a manually annotated corpus of tweets in each of the 12 languages. The multilingual dataset is created by combining all of the individual datasets. Each sample

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 357–364

Language	Train				Dev				Test			
	Total	**positive**	**negative**	**neutral**	Total	**positive**	**negative**	**neutral**	Total	**positive**	**negative**	**neutral**
am	5985	22.26%	25.87%	51.87%	1498	22.24%	25.92%	51.84%	2000	21.91%	66.88%	11.21%
dz	1652	25.26%	54.03%	20.71%	415	25.36%	53.86%	20.77%	959	34.34%	49.58%	16.08%
ha	14173	33.07%	32.27%	34.66%	2678	33.13%	33.40%	33.47%	5304	33.09%	33.17%	33.74%
ig	10193	30.26%	25.51%	44.23%	1842	30.42%	25.53%	44.05%	3683	30.36%	25.61%	44.02%
kr	3303	27.23%	34.71%	38.07%	828	27.21%	34.70%	38.09%	1027	27.10%	34.60%	38.30%
ma	5584	31.49%	29.80%	38.71%	1216	31.69%	29.63%	38.68%	2962	38.64%	28.71%	32.66%
pcm	5122	35.31%	63.29%	1.41%	1282	34.89%	63.47%	1.64%	4155	33.63%	55.99%	10.38%
pt	3064	22.23%	25.53%	52.24%	768	22.29%	25.55%	52.15%	3663	17.15%	17.89%	64.96%
sw	1811	30.22%	10.55%	59.23%	454	30.24%	10.60%	59.16%	749	29.95%	10.70%	59.36%
ts	805	47.76%	35.32%	16.92%	204	47.29%	35.47%	17.24%	255	47.64%	35.43%	16.93%
twi	3482	47.23%	37.78%	15.00%	389	47.16%	37.89%	14.95%	950	47.42%	37.20%	15.38%
yo	8523	41.56%	21.97%	36.47%	2091	42.30%	21.20%	36.51%	4516	42.48%	21.73%	35.79%
multilingual	63697	32.63%	31.57%	35.79%	13665	32.32%	31.80%	35.88%	30223	32.44%	33.78%	33.79%

Table 1: Split-wise Label Distribution of the Datasets

in the dataset consists of an ID, the text of the tweet, and a label. The label can be "positive", "neutral", or "negative". The organisers have already divided the datasets into train, dev and test splits. We notice that there is a considerable amount of variation in the label distribution between the languages. While some languages like Hausa (ha) and Darija (ma) have an almost perfectly equal representation of each of the three classes, others have a significant class imbalance in the dataset, with Nigerian Pidgin (pcm) being the biggest standout with less than 2% of the samples labelled neutral. Three other languages (Xitsonga (ts), Twi (twi), and Swahili (sw)) have one of the classes at less than 20% representation. The detailed split-wise label distribution for the datasets is given in Table 1.

Nine of the 12 languages in the dataset are in Latin script. Of the remaining three, Amharic is written in Ge'ez script, Algerian Arabic in Arabic script, and Darija in both Latin and Arabic scripts. The tweets are also code-mixed with English.

Apart from the details provided in Muhammad et al. (2023a), Muhammad et al. (2022) also described the collection and annotation process for Hauso, Igbo, Naija (Nigerian Pidgin) and Yoruba. The Amharic dataset is described by Yimam et al. (2020).

2.3 Sentiment Classification

Large, multilingual Pretrained Language Models (PLMs) such as multilingual BERT (Devlin et al., 2018) and RoBERTa (Conneau et al., 2019) and their derivatives, which have been trained on massive corpora covering 100+ languages have been shown to perform well on downstream tasks for resource-poor languages (Pires et al., 2019), including on sentiment classification in languages such as Arabic (Alammary, 2022) and Swahili (Martin et al., 2021).

For African languages, AfriBERTa (Ogueji et al., 2021) (trained on 11 languages) and AfroXLMR (Alabi et al., 2022) have shown state-of-the-art performance on various downstream tasks. This includes sentiment classification on the NaijaSenti dataset for Hausa, Igbo, Yoruba, and Nigerian Pidgin (Muhammad et al., 2022).

Ensemble learning (Zhou, 2009) is another popular method used to make classification models more robust (Kazmaier and van Vuuren, 2022). An ensemble leverages the fact that different classifiers leverage different characteristics of the input during prediction. They combine the predictions of a set of diverse models to produce a more robust output. Araque et al. (2017) show specifically that combining feature-based classifiers with deep, neural network-based classifiers can increase model performance across a variety of datasets.

We use a popular ensembling method, XGBoost (Chen and Guestrin, 2016), to combine the predictions from transformer-based classifiers as well as lightweight, feature-based classification algorithms such as SVM (Cortes and Vapnik, 1995), Random Forest (Breiman, 2001) and Logistic Regression (Cox, 1958). We also incorporate emoticon frequencies into the feature vector for our final ensemble learning model.

3 System Overview

We train an ensemble model for each language that takes in a feature vector consisting of the output of six classification models and emoticon count features for each tweet.

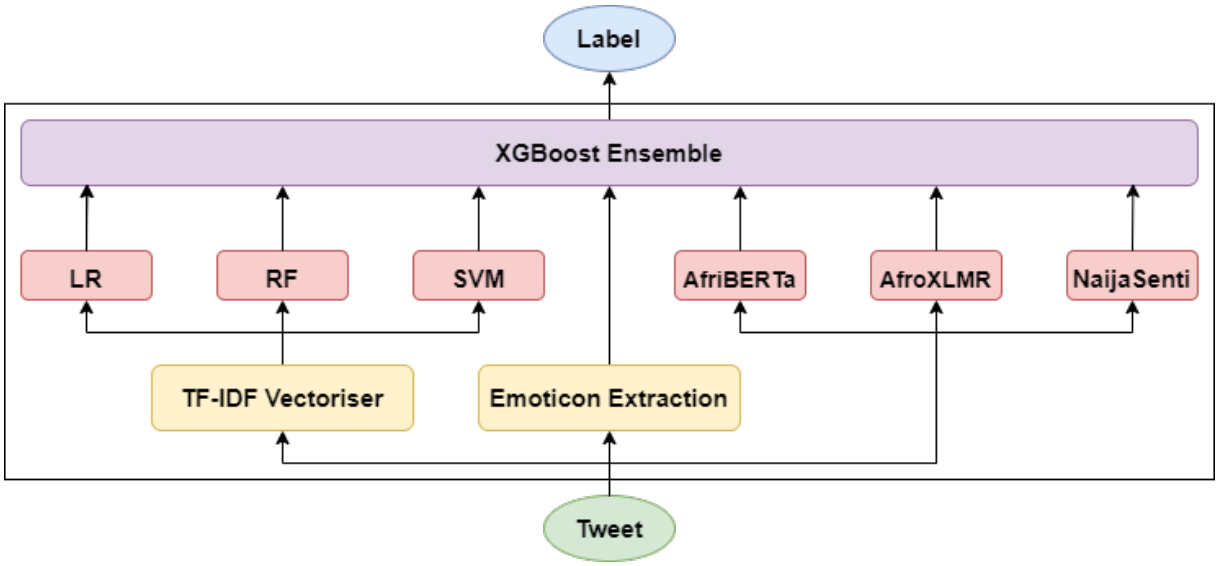

Figure 1: System Overview

3.1 Statistical models

For each language, we train three statistical models, and we obtain the probability of each class to use as features for our ensemble. We use term frequency-inverse document frequency (tf-idf) as input features to our classifiers. The classifiers we train are -

- **Logistic Regression (LR)** - we train a multi-class logistic regression (Cox, 1958) classifier using a one vs rest approach. We limit the model to 1000 iterations or convergence.

- **Support Vector Machine (SVM)** - An SVM (Cortes and Vapnik, 1995) is a binary classifier that tries to find a hyperplane that most accurately divides the training data into two classes. We train an SVM for multi-class classification using a one-vs-rest approach and a linear kernel.

- **Random Forests (RF)** - A Random Forest (Breiman, 2001) is a type of ensemble classifier that uses a large number of decision trees that are each trained using a different subset of the input features and a subset of the training dataset (with replacement), which are then combined using a bagging approach. We train our RF classifier using 100 such decision trees.

3.2 Transformer-based models

We train three different transformer models for each language and obtain the output score for each class for all the samples in the data. These scores are then used to create the feature vector for the ensemble. We use the following models -

- **AfriBERTa** - AfriBERTa (Ogueji et al., 2021) is an XLM-RoBERTa (Conneau et al., 2019) based language model that was trained on 1 GB of text data in 11 African languages, including 6 of the 12 languages in the task. It was shown to outperform mBERT and XLM-RoBERTa on various tasks, including text classification. We use the AfriBERTa-base variant that consists of 111M parameters.

- **AfroXLMR** - AfroXLMR(Alabi et al., 2022) is another language model based on XLM-RoBERTa that was trained using 17 African languages - of which 7 are a part of the task - and three high resource languages (English, French, Arabic). The model was shown to be competitive with existing models and improve zero-shot classification for unseen languages in some tasks. We use the AfroXLMR-base variant for our system.

- **AfriBERTa-NaijaSenti** - This model is a multilingual classification model based on AfriBERTa-large that achieved the best scores on the original NaijaSenti dataset (Muhammad et al., 2022), which is a part of the corpus for this task. We further fine-tune this model for each language using the respective training datasets.

3.3 Ensemble Classifier

Each model may learn different characteristics of the data towards the training task depending on the model architecture, training objective, and any pre-trained weights used. This can lead to different model capabilities, which can be leveraged by applying an ensemble over the individual model predictions. For this, we train an XGBoost (Chen and Guestrin, 2016) classifier on the train predictions of the statistical and transformer-based prediction probabilities for each sentiment class.

Emoticons are an integral part of online text-based communication and can significantly impact the tone and overall sentiment of the text. Thus, for each language, we identify the frequently occurring emoticons (present in at least 10% of the training samples). The respective frequencies of these emoticons are generated for all data samples and used as features alongside the individual model predictions in the ensemble model. Note that in case there is no emoticon that is present in at least 10% of the training samples, emoticon features are not added to the feature, and only the classification scores are used. Figure 1 illustrates the top-level view of our final system.

4 Experimental Details

For each language, we only use the train set provided by the organisers during the training of our models. The development set was used as an unseen set to compare the final performance of the various models that we trained during the competition.

All three of our transformer models are trained for five epochs on a Kaggle kernel with an Nvidia P100 GPU with 16 GB VRAM and 13 GB of RAM. We make use of the simpletransformers library (Rajapakse, 2019), which is based on the HuggingFace Transfomers library(Wolf et al., 2020).

For our statistical models, we use the implementations provided by scikit-learn (Pedregosa et al., 2011) for both tf-idf feature extraction, training our classifiers, and evaluating them. For logistic regression, maximum iterations were set to 1000 epochs, and for the random forest classifier, we set the number of decision trees to 100. Our SVM models were trained with a linear kernel. To ensure uniformity, we took the probability of each class as the feature to our final ensemble model.

The ensemble XGBoost model was trained with a learning rate of 10^{-6} for a pool of 100 estimators, i.e. trees, with a max tree depth of 32 and 10% of the columns sampled for each tree.

Language	Rank	Score
Amharic (am)	25	39.09%
Algerian Arabic (dz)	25	57.55%
Hausa (ha)	15	79.65%
Igbo (ig)	7	80.87%
Kinyarwanda (kr)	22	62.69%
Darija (ma)	27	50.68%
Nigerian Pidgin (pcm)	26	64.44%
Mozambican Portuguese (ma)	22	65.02%
Swahili (sw)	19	58.91%
Xitsonga (ts)	14	52.82%
Twi (twi)	7	66.47%
Yoruba (yo)	7	78.44%
Multilingual	12	68.84%

Table 2: Ranks and weighted-F1 scores for our system submission

5 Results and Analysis

The official rankings for the task are based on the weighted f1 scores for our systems on the test sets. Our system achieves a mixed range of results across the different tracks that we participated in. It ranks 7th on the Igbo, Twi, and Yoruba tracks, which are our best results. Apart from that, it also ranks between 12 and 15 on the multilingual, Xitsonga and Hausa tracks. The complete rankings and scores are detailed in Table 2. Apart from the final system, we also report weighted f1 scores for each of the individual models that we had trained, listed in Table 3.

We observe from the class distribution of the datasets from Table 1 and our ranks that a greater imbalance in the class distribution of the training set appears to negatively affect our system compared to other systems.

Comparing the performance of our models across languages, we found a significant correlation between the performance of AfriBERTa and NaijaSenti, with a Spearman's correlation of 0.96 (p-value < 0.0001). This is along expected lines since NaijaSenti is based on AfriBERTa. Along similar lines, AfroXLMR was seen to perform significantly differently from these two models, with a Spearman's correlation of 0.77 (p-value < 0.01) with AfriBERTa and 0.82 (p-value < 0.001) with NaijaSenti. These were even lower than its correlation with the statistical models, with correla-

Language	AfroXLMR	AfriBERTa	NaijaSenti	SVM	LR	RF	Ensemble
am	54.07%	51.57%	58.39%	33.12%	29.01%	27.85%	39.09%
dz	66.08%	47.03%	42.48%	54.34%	53.48%	54.95%	57.55%
ha	78.01%	78.46%	80.53%	73.14%	72.74%	69.89%	79.65%
ig	78.13%	79.04%	80.59%	77.69%	77.22%	74.71%	80.87%
kr	66.23%	63.25%	62.20%	57.87%	56.15%	54.31%	62.69%
ma	48.23%	41.76%	41.13%	56.59%	53.17%	46.43%	50.68%
pcm	66.70%	62.40%	68.66%	60.03%	60.79%	57.05%	64.44%
pt	68.66%	58.56%	59.48%	62.14%	61.92%	59.90%	65.02%
sw	62.00%	62.35%	60.30%	55.30%	53.67%	50.74%	58.91%
ts	39.09%	54.44%	52.87%	49.73%	49.07%	47.36%	52.82%
twi	63.00%	65.42%	64.91%	62.17%	61.90%	63.32%	66.47%
yo	69.38%	73.48%	79.29%	72.55%	72.07%	65.10%	78.44%
multilingual	68.48%	64.32%	66.74%	64.45%	64.04%	58.80%	68.84%

Table 3: Weighted F1 scores for each language and model trained for the task on the test set. The scores for the individual models were calculated after the release of the test set by us, while the scores for the ensemble (also on the same test set) were taken directly from the competition website.

tion values ranging from 0.89 to 0.92 (p-values < 0.0001).

Comparing the performance of different languages across the models, we notice that there is an observable similarity between the results in Hausa (ha), Igbo (ig), and Yoruba (yo). The Spearman's correlations for pairs of these languages ranges between 0.89 to 0.96, with p-values < 0.01. This could potentially be because these three languages have the largest amount of training data and are also part of the training corpus for all three of the transformer models.

Additionally, the pairs Kinyarwanda (kr)-Swahili (sw) and Nigerian Pidgin (pcm)-Amharic both have a correlation of 0.93 (p-value < 0.01). This could be possibly due to typological similarities between the languages, which need to be investigated further.

We also conducted an ablation study by changing the configuration of the ensemble classifier to exclude certain features after the competition once the test set labels were released. We varied the configuration to either include (+) or exclude - the emoticon features (`EMO`), and the outputs of the transformer models (`TR`) or the statistical models (`ST`). `+EMO+TR+ST` is the configuration we submitted for the competition.

We make a few observations on the basis of this set of results:

- Overall, ensembling only the transformer-based models seems to outperform all other configurations for most languages.
- Only in the case of Darija (ma), the ensemble of the statistical models outperforms those containing transformer models. This may be because of the fact that Darija is present in a mix of both Latin and Arabic scripts, while all the others are present in single script. However, further analysis is required to confirm this hypothesis.
- Since emoticon features are only generated if there is at least 1 emoticon present in at least 10% of the training data, some of the languages (am, pcm, sw) where there are not enough emoticons available perform identically whether we include that feature or not.
- Ensembles without emoticon features frequently outperform ensembles with them. Although we believe emoticons do contribute important information, the lower scores may be due to a sparsity of emoticons in the dataset, resulting in the recall being low. For example, the presence of a "positive" emoticon may be highly indicative of the overall sentiment in the tweet being positive (high precision), but its absence is not enough to ascertain that the tweet is not positive (low recall).

The weighted F1 scores for these models are reported in Table 4.

6 Additional experiments

During the development phase of the competition, we tried out several experiments that did not make

Language	+EMO			-EMO		
	+TR	+ST	+TR+ST*	+TR	+ST	+TR+ST
am	**46.51%**	38.26%	39.09%	**46.51%**	38.26%	39.09%
dz	39.92%	53.84%	**57.51%**	54.48%	54.93%	**58.37%**
ha	**81.60%**	72.17%	79.65%	**81.87%**	72.75%	79.32%
ig	79.94%	76.36%	**80.87%**	**81.34%**	77.91%	80.74%
kr	**65.84%**	56.46%	62.69%	**65.81%**	56.60%	61.64%
ma	43.91%	**53.25%**	50.68%	44.05%	**53.61%**	50.99%
pcm	**66.02%**	62.19%	64.44%	**66.02%**	62.19%	64.44%
pt	**65.19%**	60.63%	65.02%	**65.82%**	63.11%	65.28%
sw	**62.29%**	55.20%	58.91%	**62.29%**	55.20%	58.91%
ts	**53.10%**	51.03%	52.82%	**53.51%**	52.24%	52.24%
twi	63.99%	62.86%	**66.47%**	67.01%	64.15%	**66.78%**
yo	78.38%	72.90%	**78.44%**	78.38%	72.90%	**78.44%**
multilingual	**68.69%**	64.31%	68.84%	**69.21%**	63.96%	68.84%

Table 4: Ablation study of different configurations of the ensemble model. Scores reported are Weighted F1 on the test set. * - Configuration that was submitted for the competition. **Underlined** indicates the best-performing model for the language across classes. **Bold** indicates the best-performing model for a language for that class of models (based on +EMO/-EMO)

it to the final submission. This section discusses the motivation behind some of them.

Most of the tweets included in the dataset were code mixed with English. Therefore, we experimented with **replacing all emoticons with the corresponding English text** (such as replacing a smiley face with the token "smiling face"). We expected that this would help our models learn better by removing the emoticons from the vocabulary and increasing the frequency of their corresponding sentiment words. We tried doing this with AfriBERTa on Hausa and Igbo since they have the largest datasets available, and Igbo has a greater class imbalance in the training set than Hausa. However, the weighted f1 score fell from 79.88 to 77.64 for Hausa and from 80.33 to 79.12 for Igbo, so we decided not to continue with it.

We noticed that some of the models we trained were inconsistent with the neutral class, especially when there was a noticeable imbalance in the training set. To tackle this, we trained an AfriBERTa model specifically to distinguish between polar and non-polar tweets by **replacing the "positive" and "negative" labels in the dataset with a "non-neutral" label**. We expected this model to noticeably improve the classification on the neutral class and use that class score as a feature for the ensemble. However, this model showed no improvement in the f1 score for the neutral class in Igbo(0.79 f1 score for the neutral class in both cases) and scored lower in Hausa (0.74 f1 score compared to 0.78 earlier).

Finally, we also experimented with **combining the datasets for Algerian Arabic and Darija** since they are both variants of Arabic. We transliterated the datasets from Arabic script to Latin script (since Algerian Arabic was in Arabic script while Darija had a mix of both Latin and Arabic) using the Buckwalter system (Buckwalter, 2002). However, we did not notice any performance improvements in the predictions for either Algerian Arabic or Darija and decided not to pursue this experiment further.

7 Ethical Considerations

A sentiment classification model has significant potential use for online community management tasks such as forum moderation on social media platforms. If used without exhaustive evaluation and testing under different scenarios, it can cause significant damage, such as propagating any biases within the model. Even if the model is unbiased and robust, it can be used as a tool of suppression to identify and target individuals with specific viewpoints (such as their opinion of a particular organisation). Hence, developing a robust test for checking inherent biases is extremely important, as is exhaustive moderation and control over where and for what purposes such models are being deployed.

8 Conclusion and Future Work

We describe our system submission for the AfriSenti shared task at Semeval-2023. We combine the predictions of three transformer-based classifiers and three statistical ones and add emoticon frequencies to construct a feature vector for an XGBoost-based ensemble model. Though the system achieves mixed results in the rankings, we analyse the performance of each of the individual models to show a significant correlation between two of our transformer models and between several language pairs. Finally, we describe additional experiments that we conducted, which included replacing emoticons with text, combining two of the three classes to train a classifier to specifically distinguish the third class, and combining multiple datasets of similar languages to try and increase the training data. Since these experiments did not increase our model performance, we did not include them in the final system. We believe the additional investigation into combining datasets from different but related languages (such as leveraging Arabic resources for Darija and Algerian Arabic) could lead to more robust models. We would also like to investigate the correlated language pairs for linguistic and typological features that could potentially explain that observation.

References

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*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ali Saleh Alammary. 2022. Bert models for arabic text classification: A systematic review. *Applied Sciences*, 12(11).

Oscar Araque, Ignacio Corcuera-Platas, J. Fernando Sánchez-Rada, and Carlos A. Iglesias. 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246.

Leo Breiman. 2001. *Machine Learning*, 45(1):5–32.

Tim Buckwalter. 2002. Arabic transliteration. `http://www.qamus.org/transliteration.htm`.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

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.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Chayma Fourati, Hatem Haddad, Abir Messaoudi, Moez BenHajhmida, Aymen Ben Elhaj Mabrouk, and Malek Naski. 2021. Introducing a large Tunisian Arabizi dialectal dataset for sentiment analysis. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 226–230, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Jacqueline Kazmaier and Jan H. van Vuuren. 2022. The power of ensemble learning in sentiment analysis. *Expert Systems with Applications*, 187:115819.

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 Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris

Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth 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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

T. C. Rajapakse. 2019. Simple transformers. `https://github.com/ThilinaRajapakse/simpletransformers`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. 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.

Zhi-Hua Zhou. 2009. *Ensemble Learning*, pages 270–273. Springer US, Boston, MA.