Bhattacharya_Lab at SemEval-2023 Task 12: A Transformer-based Language Model for Sentiment Classification for Low Resource African Languages: Nigerian Pidgin and Yoruba

Nathaniel Hughes[1,+], **Kevan Baker**[2,+], **Aditya Singh**[1],
Aryavardhan Singh[1], **Tharalillah Dauda**[1], **and Sutanu Bhattacharya**[1,*]

[1]Department of Computer Science and Computer Information Systems,
Auburn University at Montgomery, Montgomery, AL, USA

[2]Department of Computer Science, Florida Polytechnic University, Lakeland, FL, USA

[1]{nhughes4,asingh7,asingh8,tdauda,sbhatta4}@aum.edu
[2]{kbaker5651}@floridapoly.edu

Abstract

Sentiment Analysis is an aspect of natural language processing (NLP) that has been a topic of research. While most studies focus on high-resource languages with an extensive amount of available data, the study on low-resource languages with insufficient data needs attention. To address this issue, we propose a transformer-based method for sentiment analysis for low-resources African languages, Nigerian Pidgin and Yoruba. To evaluate the effectiveness of our multilingual language models for monolingual sentiment classification, we participated in the AfriSenti SemEval shared task 2023 competition. On the official evaluation set, our group (named as Bhattacharya_Lab) **ranked 1** out of 33 participating groups in the Monolingual Sentiment Classification task (i.e., Task A) for Nigerian Pidgin (i.e., Track 4), and in the **Top 5** among 33 participating groups in the Monolingual Sentiment Classification task for Yoruba (i.e., Track 2) respectively, demonstrating the potential for our transformer-based language models to improve sentiment analysis in low-resource languages. Overall, our study highlights the importance of exploring the potential of NLP in low-resource languages and the impact of transformer-based multilingual language models in sentiment analysis for the low-resource African languages, Nigerian Pidgin and Yoruba.

1 Introduction

Detecting sentiments or emotions from language is a prominent area of research in natural language processing (NLP) (Medhat et al., 2014; Wankhade et al., 2022). Sentiment analysis (SA) deals with identifying and classifying emotions from textual data (Pang et al., 2002).

However, most of the studies focus on high resource languages like English (Ruder, 2020). As such, low resource languages such as African languages with a limited amount of data remain highly underrepresented, making the sentiment classification task for African languages is challenging (Mabokela and Schlippe, 2022). In recent years, there has been increasing interest in developing sentiment classification models for African languages (Muhammad et al., 2022; Ogueji et al., 2021).

However, the lack of high-quality labeled datasets, as well as the limited research in this area, limit the progress in this field (Muhammad et al., 2023a). To address this challenge, several research initiatives, such as the SemEval Shared Task (Muhammad et al., 2023b), have been launched to encourage the development of effective sentiment analysis models for African languages. We participated in the SemEval Shared Task 2023, specifically the Monolingual Sentiment Classification task (i.e., Task A) for two low-resource African languages: Yoruba and Nigerian Pidgin. The aim of this task is to develop NLP models capable of accurately classifying the sentiment of text in these two African languages (Track 2 and Track 4).

In this paper, we present a transformer-based language model trained on two low-resource African languages, Yoruba and Nigerian Pidgin, to correctly classify sentiments. Our model is officially ranked 1 in the monolingual sentiment classification task (Task A) for Nigerian Pidgin, and ranked 5 in the monolingual sentiment classification task (Task A) for Yoruba in the AfriSenti SemEval shared task 2023 competition, demonstrating the effectiveness of our approach

[+]These authors contributed equally to this work
[*]corresponding author

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1502–1507
July 13-14, 2023 ©2023 Association for Computational Linguistics

Language	Tweet	English translation	Sentiment
Yoruba	Ȩ kú o̧dún titun oooo!!	Happy new year!!	Positive
	Arugbo koni daa o.	Old age is not good.	Negative
Nigerian Pidgin	You gat a special place.	You have a special place.	Positive
	You still fit open mouth talk na wa oh	You're still talking it's a wonder	Negative

Table 1: Example of Tweets in Yoruba and Nigerian pidgin languages with English translation and sentiment.

in correctly classifying the sentiment of the text in Yoruba and Nigerian Pidgin by attaining the state-of-the-art performance.

2 Related Work

Recent advances in unsupervised learning of text representations lead to advancements in natural language processing problems. Pretrained word embeddings is improved by learning contextualized representations (Mikolov et al., 2013; Peters et al., 1802), and this is further improved by pretraining language models (Radford et al., 2018; Devlin et al., 2018) based on transformers (Vaswani et al., 2017). Multilingual language models, in which a single language model is pretrained on several languages without any explicit cross-lingual supervision (Conneau et al., 2019).

However, most studies focus on high-resource languages with relatively large amounts of data. As such, low resource languages such as African languages with a limited amount of data remain highly underrepresented (Muhammad et al., 2022; Ogueji et al., 2021). (Ezeani et al., 2018; Ogueji and Ahia, 2019; Alabi et al., 2019; Dossou and Sabry, 2021) show promising results in monolingual cases using pretrained embeddings for African languages, however, the models are static and trained on a specific language. Recent studies on multilingual language model (Devlin et al., 2018; Liu et al., 2019; Conneau et al., 2019) show superior performance, particularly for high-resource languages. While such models use a large training data, there is a need for increased representation of low-resource African languages in modern multilingual language model (Muhammad et al., 2022).

Inspired by a recent study (Ogueji et al., 2021) using a relatively small amount of data, this work focuses on pretraining a multilingual language model solely on low-resource languages without any transfer from higher-resource languages. We also take leverage of using a recently released dataset, mostly manually annotated, of low resource languages originating from Nigeria (Muhammad et al., 2022).

3 Methodology

3.1 Languages - Nigerian Pidgin and Yoruba

We focus on two African languages, Nigerian Pidgin and Yoruba. Tables 1 and 2 provide the details about both languages. In particular, Table 1 presents sample tweets in both languages along with their English translation. Three classes of sentiments are considered: (a) positive sentiment, i.e., a tweet characterized by emotions such as happiness, satisfaction, contentment, joy, excitement, and optimism; (b) negative sentiment, i.e., a tweet characterized by emotions such as anger, frustration, disappointment, sadness, and pessimism; and (c) neutral, i.e., a tweet characterized by a sense of indifference or neutrality. Examples of positive and negative sentiments in both languages are included in Table 1. Moreover, Table 2 provides additional information about both languages. While Nigerian Pidgin belongs to the English Creole family of language with around 75M speakers, Yoruba belongs to the Niger-Congo family with around 42M speakers.

Language	Family	Speakers
Nigerian Pidgin	English Creole	75M
Yoruba	Niger-Congo	42M

Table 2: Information of representative languages

3.2 Experimental Setup

3.2.1 Sentiment Classification Model

We train a transformer (Vaswani et al., 2017) using the standard masked language modeling objective (Devlin et al., 2018) without next-sentence prediction, which is the same approach used in XLM-R (Conneau et al., 2020). We train on text data that includes both languages and sample batches from both languages, ensuring that the model does not encounter the same language in consecutive batches. To tokenize the raw text data, we use subword tokenization with SentencePiece (Kudo and Richardson, 2018), which is trained with a unigram language model (Kudo, 2018). We follow the sampling method described by (Ogueji et al., 2021) with a parameter alpha of 0.3 to sample training sentences. The Huggingface Transformers library (Wolf et al., 2020) is used to train the model. We use a maximum sequence length of 256 and train the model for 25 epochs with a batch size of 32, warm-up steps of 100, and a default learning rate of 5e-5. Moreover, we optimize the models using AdamW (Loshchilov and Hutter, 2017). After fine-tuning, our model has 10 layers, 6 attention heads, 768 hidden units, and 3072 feed-forward size.

3.2.2 Data sets

Training data sets: To train our model, we use the training set used by (Ogueji et al., 2021) along with AfriSenti training data (Muhammad et al., 2023a). Tables 3 and 4 provide details of both data sets. *Development data set:* As shown in Table 3, we use AfriSenti development data. Specifically, it contains 1282 tweets in Nigerian pidgin and 2091 tweets in Yoruba, respectively. According to AfriSenti, the proportion of tweets in each label (positive, negative, and neutral) varies significantly for Nigerian Pidgin than that of Yoruba. *Test data set:* We evaluate the performance of our model on the SemEval official evaluation set. It is worth mentioning that we participated in the SemEval Shared Task 2023, specifically the Monolingual Sentiment Classification task (i.e., Task A) for Yoruba (i.e., Track 2) and Nigerian Pidgin (i.e., Track 4).

Datasets	Nigerian Pidgin	Yoruba
Train	5122	4155
Dev	1282	2091
Test	4155	4516

Table 3: AfriSenti data sets split information of representative languages (Muhammad et al., 2023a). The number of Tweets in each category is reported.

Language	# sentences	Size (GB)
Nigerian Pidgin	161,842	0.05
Yoruba	149,147	0.03

Table 4: Training data sets information of representative languages (Ogueji et al., 2021).

3.2.3 Performance Evaluation

We use precision, recall, and weighted F1 score to evaluate the performance of our model. A high precision score indicates that the model is making very few false positive predictions, while a high recall score indicates that the model is correctly identifying a high proportion of actual positive instances. The F1 score is a harmonic mean of precision and recall, and the weighted F1 score is a variation of F1 score that accounts for class imbalance in the dataset. A high F1 score indicates the better performance.

4 Results

4.1 Performance on Development Set

Method	Nigerian Pidgin	Yoruba
Precision	0.762	0.798
Recall	0.760	0.799
F1 score	0.760	0.799

Table 5: Performance of our model on the development set, containing 1282 Nigerian Pidgin and 2091 Yoruba tweets. Here, weighted F1 score is reported.

Table 5 represents the performance of our model on the AfriSenti development data set, containing 1282 and 2091 tweets for Nigerian Pidgin and Yoruba, respectively. Our model consistently achieves high performance, with weighted F1 scores of 76% and 79.9% for Nigerian Pidgin and Yoruba, respectively. We observe similar trends by using Precision and Recall evaluation metrics. It is worth noting that the proportion of tweets in each label (positive, negative, and neutral) varies significantly for Nigerian Pidgin, whereas this is not as significant for Yoruba according to AfriSenti, illustrating that our model works well in both cases.

4.2 Performance on Nigerian Pidgin Test Set

Figure 1 shows the head-to-head performance comparison of our method (named Bhattacharya_lab) against the competitive participating methods where the ranking was officially released by the AfriSenti SemEval organizer. The test set contains 4155 tweets. We **ranked 1** out of 33 participating groups by attaining a weighted F1 score of 75.96%, illustrating the superior performance of our transformer-based model in sentiment analysis for the low-resource African language Nigerian Pidgin. In particular, the performance gap between

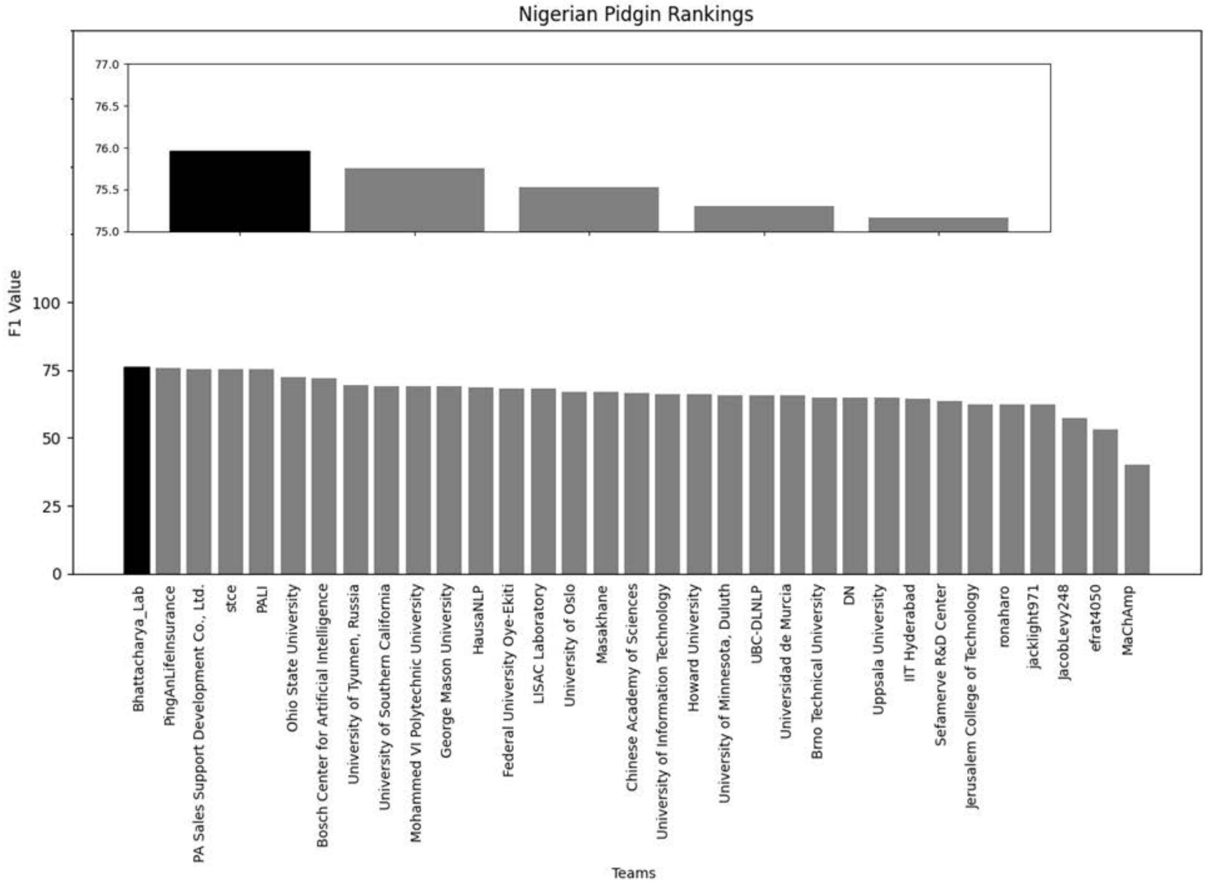

Figure 1: A head-to-head performance comparison of our method (named Bhattacharya_lab) in Black against the competitive participating methods in Grey where the ranking was officially released by the AfriSenti SemEval organizer (Task A, Track 4) over the Nigerian Pidgin dataset of 4155 tweets. Our method officially ranked 1 out of 33 participating groups. The inset figure illustrates the performance of top five groups.

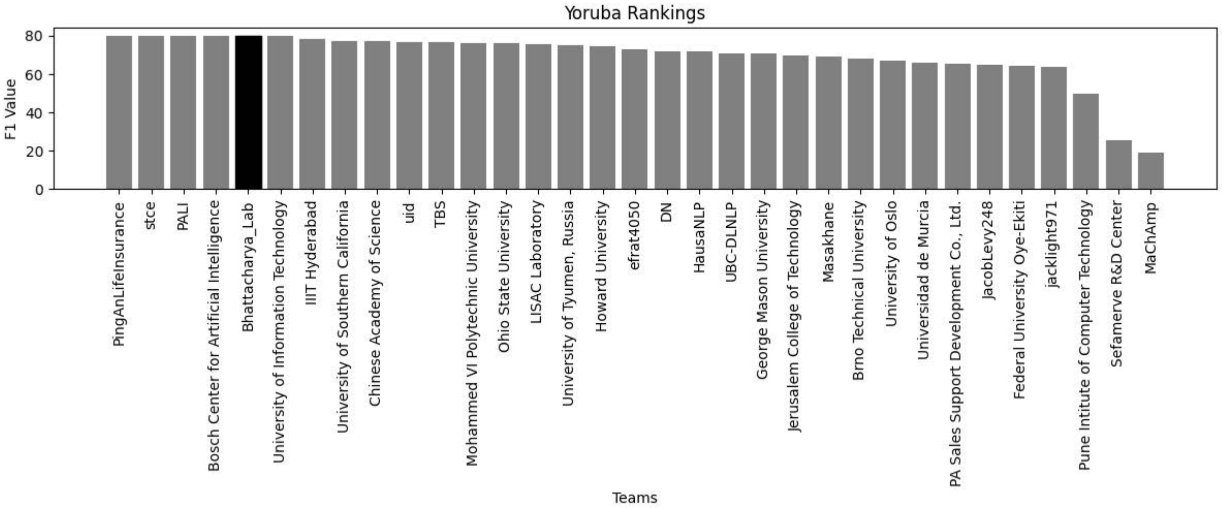

Figure 2: A head-to-head performance comparison of our method (named Bhattacharya_lab) in Black against the competitive participating methods in Grey where the ranking was officially released by the AfriSenti SemEval organizer (Task A, Track 2) over the Yoruba dataset of 4516 tweets. Our method officially ranked 5 out of 33 participating groups.

the top-ranked method (this work) and the top-5 ranked method (named Pali) is 0.8% F1 score. It is worth mentioning the official method description of other participating methods are yet not published, therefore, we cannot compare top-ranked methods based on their methodologies.

4.3 Performance on Yoruba Test Set

Figure 2 shows the head-to-head performance comparison of our method (named Bhattacharya_lab) against the competitive participating groups where the ranking was officially released by the AfriSenti SemEval organizer. On this dataset of 4516 tweets, we ranked in the **Top 5** out of 33 groups with a weighted F1 score of 79.86%. In particular, the performance of our work is significantly superior to the majority of the participating groups, we perform comparably to the top1-ranked group (79.86% vs 80.16%). That is our performance in the Yoruba language is consistent with our performance in the Nigerian Pidgin language, illustrating the impact of our transformer-based multilingual language models in sentiment analysis for the low-resource African languages, Nigerian Pidgin and Yoruba.

5 Conclusion

The study of low-resource languages with insufficient data requires attention since most of the research concentrates on high-resource languages with a large quantity of available data. We train a transformer-based model for sentiment analysis for two low-resource African languages namely Yoruba and Nigerian Pidgin. On the official evaluation set, our model consistently ranked in the Top 5 by outperforming most of the participating groups in the AfriSenti SemEval shared task 2023 competition, illustrating the superior performance of our transformer-based model over existing works.

Our contribution to this task is significant, as it demonstrates the potential for NLP techniques to be applied to low-resource languages and improve sentiment analysis in these languages. In future work, we plan to investigate the effectiveness of other pre-trained models, as well as explore the use of more advanced techniques such as multi-task learning and transfer learning. Overall, we hope that our work will encourage further research in the field of NLP for low-resource languages and contribute to the development of language technologies that can benefit underrepresented communities.

References

Jesujoba O Alabi, Kwabena Amponsah-Kaakyire, David I Adelani, and Cristina Espana-Bonet. 2019. Massive vs. curated word embeddings for low-resourced languages. the case of yor\ub\'a and twi. *arXiv preprint arXiv:1912.02481*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

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. *arXiv preprint arXiv:1911.02116*.

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

Bonaventure FP Dossou and Mohammed Sabry. 2021. Afrivec: Word embedding models for african languages. case study of fon and nobiin. *arXiv preprint arXiv:2103.05132*.

Ignatius Ezeani, Ikechukwu Onyenwe, and Mark Hepple. 2018. Transferred embeddings for igbo similarity, analogy, and diacritic restoration tasks. In *Proceedings of the Third Workshop on Semantic Deep Learning*, pages 30–38.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam.

Ronny Mabokela and Tim Schlippe. 2022. A sentiment corpus for south african under-resourced languages in a multilingual context. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 70–77.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

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 and Orevaoghene Ahia. 2019. Pidginunmt: Unsupervised neural machine translation from west african pidgin to english. *arXiv preprint arXiv:1912.03444*.

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.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 1802. Deep contextualized word representations. corr abs/1802.05365 (2018). *arXiv preprint arXiv:1802.05365*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Sebastian Ruder. 2020. Why you should do nlp beyond english. *Sebastian Ruder https://ruder. io/nlp-beyond-english*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 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.

A Appendix: Precision, Recall, and weighted F1 score

Precision measures the proportion of true positive predictions out of all predicted positive instances.

Precision = True Positives / (True Positives + False Positives)

Recall measures the proportion of true positive predictions out of all actual positive instances.

Recall = True Positives / (True Positives + False Negatives)

The F1 score is a harmonic mean of precision and recall, and the weighted F1 score is a variation of F1 score that accounts for class imbalance in the dataset.

weighted F1 score = (sum of F1 scores for each class * number of instances in each class) / total number of instances