TRANSFER LEARNING ON KINYARWANDA TWEETS SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained models available on platform such as Hugging Face have become a valuable resource for machine learning community, particularly for natural language processing task. In this study, we evaluated the performance of Kinyarwanda and English pretrained models for sentiment analysis of Kinyarwanda tweets through transfer learning using Hugging Face pretrained models and Trainer for implementation. We have found that English pretrained models for translated Kinyarwanda tweets dataset using Google translate out performed Kinyarwanda pretrained models.

1 INTRODUCTION

Kinyarwanda is a unique and beautiful language spoken in Rwanda, with a rich cultural heritage and a growing presence in the digital world. Despite its significance, Kinyarwanda has been largely overlooked in the field of natural language processing (NLP) due to a lack of available data ready for machine learning task(see Orife et al. (2020) for more information). In an effort to address this gap, we conducted a comparative analysis of open source state-of-the-art pretrained models on Kinyarwanda labeled tweets dataset, which are both available on Hugging Face. Our findings shed light on the potential of NLP for Kinyarwanda and provide valuable insights for researchers and developers interested in this exciting and underrepresented language.

2 RELATED WORK

Sentiment analysis on African languages, including Kinyarwanda, is a growing area of research, although the number of studies is much lower compared to highly represented languages such as English. Furthermore, many existing studies on African languages are unpublished or published under closed access (see Mesthrie (1995) as cited in Orife et al. (2020) for more information). Muhammad et al. (2022) conducted sentiment analysis on four Nigerian languages by collecting, filtering, processing, and labeling the dataset, and then applying transfer learning using pretrained models. Kwaik et al. (2020) employed transfer learning on pretrained models for sentiment analysis on Arabic tweets dataset. In a different study, Bataa & Wu (2019) investigated transfer learning for sentiment analysis on Japanese language using the Rakuten product review and Yahoo movie review datasets. These studies demonstrate the potential of transfer learning and pretrained models for sentiment analysis in underrepresented languages.

3 METHODOLOGY

Muhammad et al. (2023) collected a dataset of more than 110,000 annotated tweets covering 14 underrepresented African languages, including Kinyarwanda. We used the dataset, merging the train and validation sets into one and preprocessing the text by removing stopwords, URLs, and emojis, and lowercasing both the train and test sets. For transfer learning to the sentiment analysis task, we used Kinyarwanda pretrained models trained for mask task. In addition, we performed transfer learning on English pretrained models for tweet classification using translated dataset to English using the Google Translate API. For further exploration we also did preprocessing of translated tweets by removing stopwords(except negative stopwords), punctuation and lastly stemmed. We

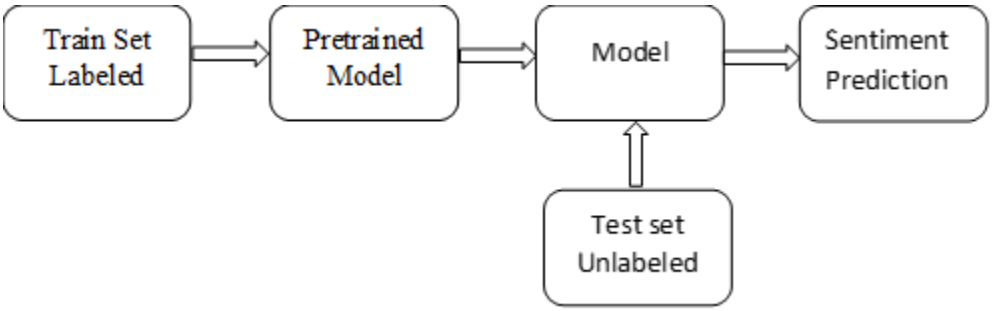

Figure 1: Methodology Flowchart.

evaluated the model's performance by comparing the F1 score of the predicted test set with the actual labeled set, using the trainer framework.

4 RESULTS

We conducted transfer learning and fine-tuning of the pretrained models using a standard tutorial as a guide (Amy, 2023). The results show that KinyaBERT-large outperformed xlm-roberta-base-finetuned-kinyarwanda while using Kinyarwanda preprocessed tweets.

Model	**Test Dataset**	**F1 Score**
bert-base-cased	translated test tweets	0.661578923
bert-base-cased	preprocessed translated test tweets	0.625459922
distilbert-base-uncased-finetuned-sst-2-english	translated test tweets	0.642492736
twitter-xlm-roberta-base-sentiment	translated test tweets	0.655731011
twitter-roberta-base-sentiment-latest	translated test tweets	0.686036459
KinyaBERT-large	Preprocessed Kinyarwanda Test Tweets	0.644029075
xlm-roberta-base-finetuned-kinyarwanda	Preprocessed Kinyarwanda Test Tweets	0.598704083

Table 1: F1 scores of various transformer models on different test datasets.

Additionally, the twitter-roberta-base-sentiment-latest model exhibited superior performance compared to other pretrained models for Kinyarwanda preprocessed translation to English. In terms of overall performance, the English pretrained models demonstrated better results than Kinyarwanda pretrained models for the sentiment analysis task through transfer learning.

5 CONCLUSION

Based on the comparative analysis of the pretrained models on Kinyarwanda tweets sentiment transfer learning task, we can conclude that the English language pretrained models outperform the Kinyarwanda masked pretrained models. This indicates the importance of having more labeled data and pretrained models in underrepresented African languages like Kinyarwanda. Among the models we experimented with, the twitter-roberta-base-sentiment-latest model performed the best with an F1 score of 0.686, closely followed by the bert-base-cased model with an F1 score of 0.661 on the test set. However, it is worth noting that the performance of the models could be improved with more fine-tuning and optimization. Overall, our study highlights the need for more research and development of NLP tools and resources for underrepresented African languages, including Kinyarwanda.

URM STATEMENT

We acknowledge this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane Mboup, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F.P. Dossou, Kelechi Ogueji, Thierno Ibrahima Diop, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 10 2021. ISSN 2307387X. doi: 10.1162/TACL_A_00416/107614/MASAKHANER-NAMED-ENTITY-RECOGNITION-FOR-AFRICAN. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00416/107614/MasakhaNER-Named-Entity-Recognition-for-African.

@GrabNGoInfo Amy. Transfer learning for text classification using hugging face transformers trainer, 2023. URL https://medium.com/grabngoinfo/transfer-learning-for-text-classification-using-hugging-face-transformers-trainer-

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. *2022 Language Resources and Evaluation Conference, LREC 2022*, pp. 258–266, 4 2021. URL https://arxiv.org/abs/2104.12250v2.

Enkhbold Bataa and Joshua Wu. An investigation of transfer learning-based sentiment analysis in japanese. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 4652–4657, 5 2019. doi: 10.18653/v1/p19-1458. URL https://arxiv.org/abs/1905.09642v3.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, 10 2018. URL https://arxiv.org/abs/1810.04805v2.

Jean Paul Ishimwe. jean-paul/kinyabert-large · hugging face, 2021. URL https://huggingface.co/jean-paul/KinyaBERT-large.

Kathrein Abu Kwaik, Stergios Chatzikyriakidis, Simon Dobnik, Motaz Saad, and Richard Johansson. An arabic tweets sentiment analysis dataset (atsad) using distant supervision and self training, 2020. URL https://aclanthology.org/2020.osact-1.1.

Rajend Mesthrie. Language and social history: Studies in south african sociolinguistics, 1995. URL http://books.google.com/books?id=aIivedw-oZYC&pgis=1.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. 1 2022. URL https://arxiv.org/abs/2201.08277.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis Davis, 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. Afrisenti: A twitter sentiment analysis benchmark for african languages. 2 2023. URL http://arxiv.org/abs/2302.08956.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. Masakhane – machine translation for africa. 3 2020. doi: 10.18653/v1/P19-1310. URL https://arxiv.org/abs/2003.11529v1.

Andre Niyongabo Rubungo. Kkltk: Kinyarwanda and kirundi languages toolkit, 2020. URL https://github.com/Andrews2017/kkltk.git.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. 10 2019. URL https://arxiv.org/abs/1910.01108v4.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

A Appendix

A.1 Acknowledgement

We used several pre-trained models for sentiment analysis, including BERT-base-cased, DistilBERT-base-uncased-finetuned-SST-2-English, and Twitter-RoBERTa-base-sentiment-latest, as well as two Kinyarwanda-specific models, KinyaBERT-large and XLM-RoBERTa-base-finetuned-Kinyarwanda see Table 1. **Note**: Pre-trained models and dataset[1] were sourced from the Hugging Face Transformers library (Wolf et al., 2019). Model details and sources: BERT-base-cased (Devlin et al., 2018)[2], DistilBERT-base-uncased-finetuned-SST-2-English (Sanh et al., 2019)[3], Twitter-RoBERTa-base-sentiment-latest[4] and Twitter-XLM-RoBERTa-base-sentiment (Barbieri et al., 2021)[5], KinyaBERT-large (Ishimwe, 2021)[6], XLM-RoBERTa-base-finetuned-Kinyarwanda (Adelani et al., 2021)[7].Lastly we have used Kinyarwanda stopwords (Rubungo, 2020)[8].

[1]https://huggingface.co/datasets/shmuhammad/AfriSenti-twitter-sentiment
[2]https://huggingface.co/bert-base-cased
[3]https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english
[4]https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest
[5]https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment
[6]https://huggingface.co/jean-paul/KinyaBERT-large
[7]https://huggingface.co/Davlan/xlm-roberta-base-finetuned-kinyarwanda
[8]https://github.com/Andrews2017/kkltk