STANCE DETECTION FOR A LOW-RESOURCE LANGUAGE

Natural Language Processing 2

Ahmad Mustapha Wali, Alexandru Ștefan Ghiță, Lara Tomeh, Radu Puținelu
University of Bucharest

ABSTRACT

Stance detection is defined as automatically determining whether the author of a piece of text supports or opposes a given target. For a variety of social and political applications where sentiment analysis may be inadequate, stance detection on social media is an emerging opinion mining paradigm. For this project, we experimented with stance detection for Zulu, a low-resource language, using several BERT-based models to observe how it compares to a previous approach to the same task that was solved using ULMFiT.

INTRODUCTION

Stance detection on social media is being used for a range of social and political applications where sentiment analysis may not be sufficient. A speaker's stance is an indication of their perspective and assessment on a certain topic. Analysis studies gauging public opinion on social media, particularly on political and social matters, heavily rely on stance detection. These topics tend to be divisive in nature, with people expressing conflicting viewpoints on matters like politics, religion, the legality of abortion, etc.

As sentiment analysis is not always enough in certain circumstances, stance detection offers a more complex and thorough knowledge of the author's viewpoint. The goal of this study was to apply stance detection to Zulu, a language with limited data and computational tools. To evaluate the efficacy of this strategy, a number of BERT-based models were tested and compared to a previous solution that used Universal Language Model Fine-tuning (ULMFiT). The aim was to investigate how these various models worked and which one was most successful in solving the problem of stance identification for Zulu.

DATA, MODEL & METHODOLOGY

Datasets:

1. **The Zulu Stance dataset**

The Zulu Stance dataset contains text samples in Zulu with annotated stance labels. It is made up of stance labels that represent the stance expressed in five different targets. The dataset consists of 1343 samples and 5 targets and is a subset of the SemEval-2016 Stance Dataset translated to Zulu using Google's Translate API. [1]

2. **The Tweet Eval Stance dataset**

The SemEval-2016 Stance Dataset is a collection of English tweets with stance annotations on abortion, climate change, atheism, Hillary Clinton, and feminism. We tried expanding the Zulu Stance dataset with 200 random samples from each topic from the SemEval-2016 Stance Dataset for the project to see how it affects the models. However, only the train split of the dataset was augmented. [2]

Model & Methodology:

We used numerous pre-trained BERT-based models for the project. The models were either pre-trained in Zulu or another language in the same family.

1. **XLM-R**

XLM-R was developed by modifying the XLM-R-large model for masked language models (MLM) use on 17 African languages (Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Oromo, Naija, Kinyarwanda, Kirundi, Shona, Somali, Sesotho, Swahili, isiXhosa, Yoruba, and isiZulu), which encompass (Arabic, French, and English). [3]

2. **Multilingual BERT**

A masked language modeling (MLM) objective was used to train a model on the top 104 languages with the largest Wikipedia. It was first introduced in this work and made available in this repository. This model was not pretrained on Zulu. [4]

Model	Score (Zulu) Train / Test	Score (Zulu + English) Train / Test
afro-xlmr-base [3]	0.54935 0.54478	0.39561 0.51493
bert-base-multilingual-cased [4]	0.49069 0.49254	0.61958 **0.58209**
afriberta_large [5]	0.46276 0.48507	0.59643 0.55224
afroxlmr-large-ner [6]	0.49907 0.52985	0.60752 0.52985
xlm-roberta-base-finetuned-zulu [3]	0.55970 **0.57821**	0.59595 0.55970
ULMFiT (Dlamini et. al.) [1]	N/A 0.4861	N/A 0.5448

CONCLUSIONS, LIMITATIONS & FUTURE WORK:

While some of the transformer models fared worse than the ULMFiT model, BERT-based models did better overall, even when not trained on the target language. Additionally, supplementing the dataset with English considerably improved performance for certain models. Because of the scarcity of training data, the performance of all models was constrained. Experimenting with a much larger dataset is a proposal for future work.

References:

1. Dlamini, Gcinizwe, Bekkouch, Imad Eddine Ibrahim, Khan, Adil, Derczynski, and Leon. 2022. *Bridging the Domain Gap for Stance Detection for the Zulu language.*
2. Barbieri, Francesco, Camacho-Collados, Jose, Espinosa-Anke, Luis, Neves, and Leonardo. 2022. *TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification.*
3. Alabi, Jesujoba O, Adelani, David Ifeoluwa, Mosbach, Marius, Klakow, and Dietrich. 2022. *Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning.* (Davlan/xlm-roberta-base-finetuned-zulu&Davlan/afro-xlmr-base)
4. Jacob Devlin and Ming-Wei Chang and Kenton Lee and Kristina Toutanova. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language.* (bert-base-multilingual-cased)
5. Ogueji, Kelechi, Zhu, Yuxin, Lin, and Jimmy. 2021. *Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages.* (castorini/afriberta_large)
6. David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Oluwadara Alabi, Shamsuddeen Hassan Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing K, Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris C, Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine W. Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen R. Gwadabe, Tosin P. Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius M Ezeani, Chiamaka Ijeoma Chukwuneke, Mofetoluwa Adeyemi, Gilles Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition. (masakhane/afroxlmr-large-ner-masakhaner-1.0_2.0)