AfriSenti-SemEval Shared Task 12

Sentiment Analysis for 15 Low-resource African Languages using Twitter Dataset

HausaNLP, MasaKhaneNLP, ICT4D, Google Research, UP Data Science for Social Impact

🌍 https://afrisenti-semeval.github.io   ✉ afrisenti-semeval-organizers@googlegroups.com

Introduction

Due to the widespread use of the Internet and social media platforms, most languages are becoming digitally available. This allows for various artificial intelligence (AI) applications that enable tasks such as sentiment analysis, machine translation and hateful content detection. According to UNESCO (2003), 30% of all living languages, are African languages.

However, most of these languages do not have curated datasets for developing such AI applications. Recently, various individual and funded initiatives, such as the Lacuna Fund, have set out to reverse this trend and create such datasets for African languages. However, research is required to determine both the suitability of current natural language processing (NLP) techniques and the development of novel techniques to maximize the applications of such datasets.

Why the Shared Task?

Previous SemEval shared tasks on sentiment analysis include Mohammad, Saif M et al., (2018), Nakov et al., (2016), Pontiki et al., Ghosh et al., (2015), (2014), and so on. However, none of these tasks included African languages. We believe SemEval is the right venue, due to its popularity and widespread acceptance, to carry out shared tasks for African languages to strengthen their further development.

Why Participate?

1. Promote NLP research involving African languages,

2. Opportunity to write and submit a system-description paper to SemEval2023 workshop to be co-located with a major NLP conference. The paper will be part of the SemEval-2023 workshop proceeding

3. Stand a chance to win award.

Task Overview

The Shared Task is based on a collection of Twitter datasets in 15 African languages for sentiment classification. It consists of three sub-tasks. Participants can select one or more tasks depending on their preference.

Task-A: Monolingual Sentiment Classification
Given training data in a target language, determine the polarity of a tweet in the target language (positive, negative, or neutral). If a tweet is conveying both a positive and negative sentiment, whichever is the stronger sentiment should be chosen.

Task-B: Multilingual Sentiment Classification
Given a combined training data from 10 African languages, determine the polarity of a tweet in the target language (positive, negative, or neutral)

Task-C: Zero-Shot Sentiment Classification
Given unlabeled tweets in two African languages (Tigrinya and Kinyarwanda), leverage any or all of the available training datasets in Subtasks 1 and 2 to determine the sentiment of a tweet in the two target languages is positive, negative, or neutral.

AfriSenti Languages

- ✓ Algerian Arabic
- ✓ Amharic
- ✓ Hausa
- ✓ Igbo
- ✓ isiZulu
- ✓ Nigerian Pidgin
- ✓ Setswana
- ✓ Swahili
- ✓ Tigrinya
- ✓ Tunisian Arabizi
- ✓ Yoruba
- ✓ Mozam. Portuguese
- ✓ Tsonga
- ✓ Twi
- ✓ Kinyarwanda
- ✓ ...

**We are expecting to add more languages.

Award and Leagues

A total of **$5,000** will be awarded to the best three teams of each league of the AfriSenti-SemEval task.

Categories

1. **African League**: To encourage African participation, this league is for only Africans.
2. **Women League**: To encourage women participation, this league is dedicated to female only.
3. **Worldwide League**: Be a participant from any country.

How to Participate

Participants can work on any or all of the sub-tasks in the shared tasks. :

A The participants can form a team with multiple people, or a single person team is okay.

B The participants can experiment with the training data to develop models. Usage of any external data or resources is allowed and highly encouraged. This process can run until the evaluation period.

C Organisers will release the test set containing instances without the labels. The participants will use their developed models to predict the labels for the instances. These predictions will be compared against the ground truth labels of the test data and the teams will be ranked on a leaderboard according to the performance score.

D Each team is encouraged to write a system description paper describing their submission system. Accepted papers will be published as part of the proceedings of the SemEval 2023 Workshop

Competition starts 1st September, 2022. For more information, visit: afrisenti-semeval.github.io/

Organizers

1. Shamsuddeen Muhammad
2. Ibrahim Sa'id Ahmad
   Affl. Bayero University, Kano–Nigeria
3. Idris Abdulmumin
   Affl. Ahmadu Bello University, Zaria–Nigeria
4. David Ifeoluwa Adelani
   Affl. Saarland University–Germany
5. Sebastian Ruder
   Affl. Google Research
6. Seid Muhie Yimam
   Affl. Universitat Hamburg, Hamburg–Germany
7. Abinew Ali Ayele
   Affl. Bahir Dar University–Ethiopia
8. Nedjma Ousidhoum
   Affl. University of Cambridge–United Kingdom
9. Vukosi Marivate
   Affl. University of Pretoria–South Africa
10. Saif M. Muhammad
    Affl. National Research Council–Canada
11. Meriem Beloucif
    Affl. Uppsala University–Sweden
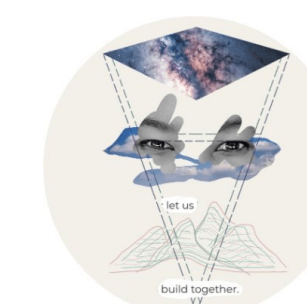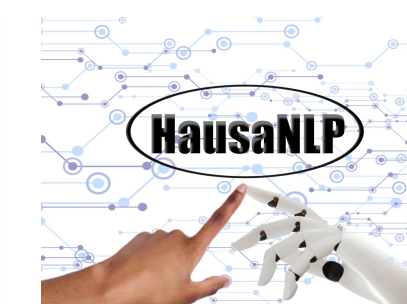
References

[1] Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris C. Emezue, Anuoluwapo Aremu, Saheed Abdul, and Pavel B. Brazdil. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *ArXiv*, abs/2201.08277, 2022.

[2] Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ali Ayele, and Chris Biemann. Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *COLING*, 2020.

Acknowledgements