

Challenges of Low-Resource Natural Language Processing: A Focus on Sentiment Analysis and Hate Speech Detection in Amharic



Dr. Seid Muhie Yimam
House of Computing and Data Science (HCDS)
Universität Hamburg, Germany

Cambridge NLP Seminars

19th May 2023

Slides adapted from my AfricaNLP 2023 talk

outline

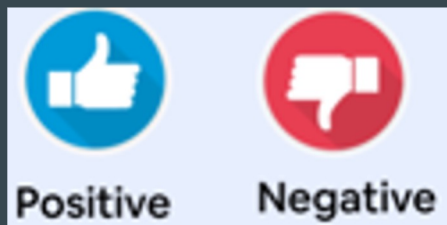
- Low-resource-ness

- Tasks
- Initiatives
- Challenges



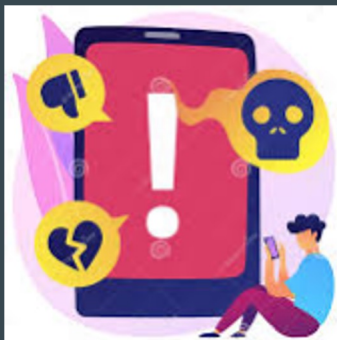
- Sentiment Analysis

- Annotation and model
- Challenges



- Hate speech detection

- Annotation and model
- Challenges



- Conclusion

NLP research focuses

Kalika Bali, Monojit Choudhury, Sunayana Sitaram, Vivek Seshadri (2019) ELLORA: Enabling Low Resource Languages with Technology

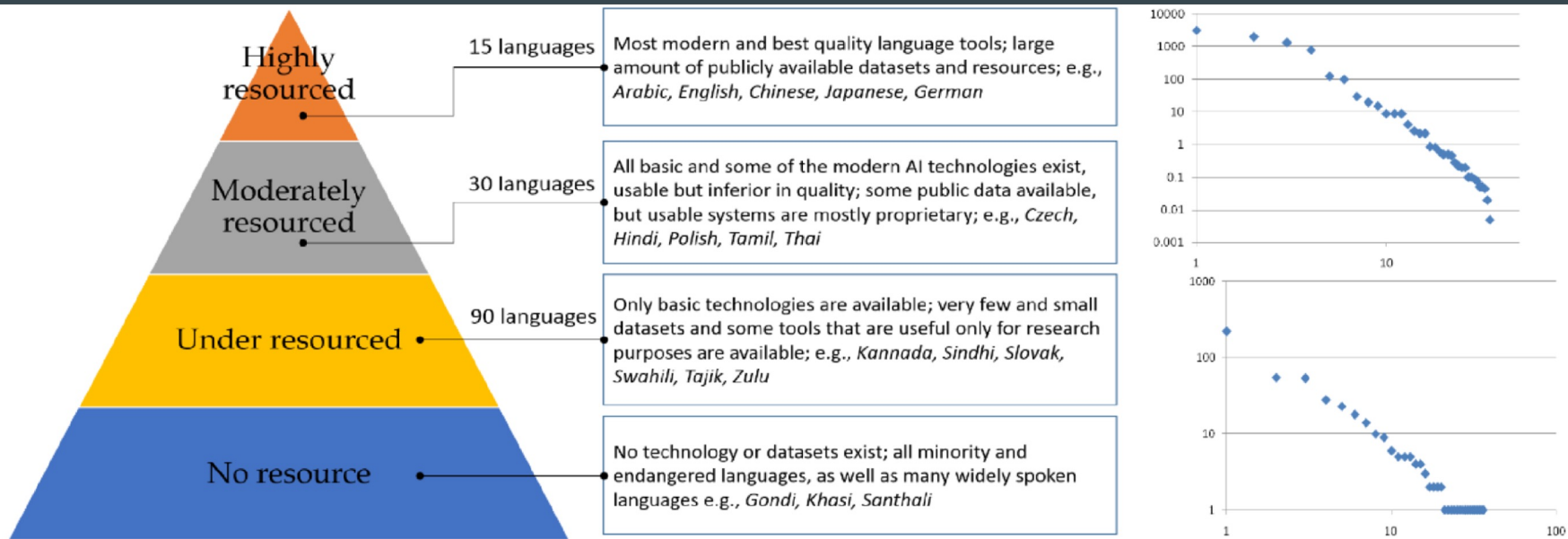
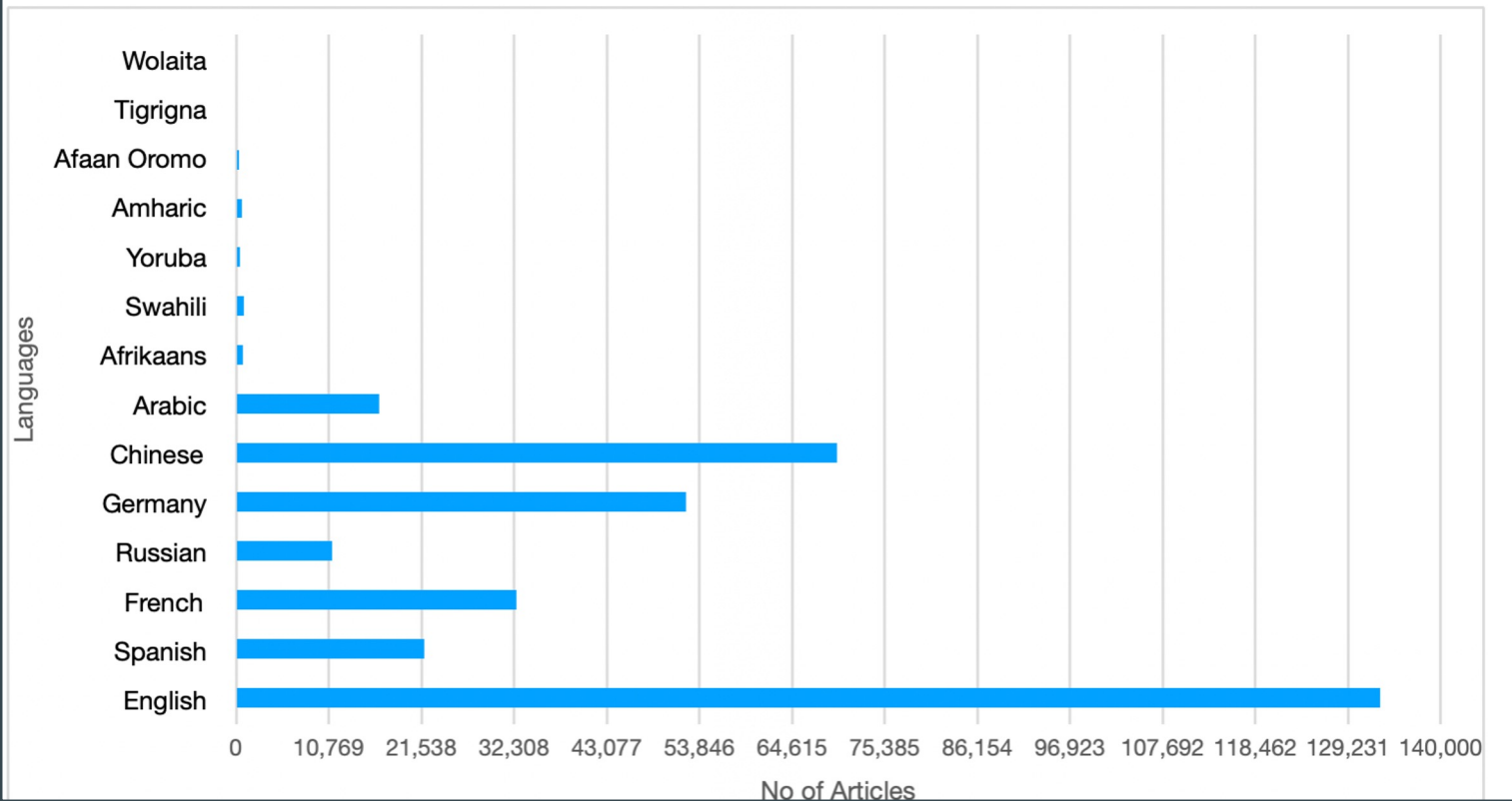


Figure 1: Classification of languages according to the availability of language technology, tools and resources (left) based on the power-law distribution of the resources across the languages of the world (right).

Research focus in Africa

- Most African countries allocate less than **1%** of their budget for Research
- According to the Ethiopian Technology and Innovation Institute report of 2017, Ethiopia allocates **0.27%** of its budget for research



ACL Anthology paper search

Tonja et al. (2023): Natural Language Processing in Ethiopian Languages: Current State, Challenges, and Opportunities

Part-of-speech	37500
"Part-of-speech"	25000
Amharic Part-of-speech	293
"Amharic Part-of-speech"	10
Tigrinya Part-of-speech	114
"Tigrinya Part-of-speech"	1
Oromo Part-of-speech	119
"Oromo Part-of-speech"	0
speech recognition 2019	27300
"speech recognition"	11700
Amharic speech recognition	221
"Amharic speech recognition"	9
Tigrinya speech recognition	127
"Tigrinya speech recognition"	6
Oromo speech recognition	90
"Oromo speech recognition"	0

ACL
Anthology
Phrase search

Phrase	Count
Part-of-speech	54,100
"Part-of-speech"	33,500
Amharic Part-of-speech	397
"Amharic Part-of-speech"	19
Tigrinya Part-of-speech	132
"Tigrinya Part-of-speech"	
Oromo Part-of-speech	137
"Oromo Part-of-speech"	0
speech recognition 2023	35,100
"<u>s</u>peech recognition"	15,600
Amharic speech recognition	313
"Amharic speech recognition"	59
Tigrinya speech recognition	116
"Tigrinya speech recognition"	7
Oromo speech recognition	127
"Oromo speech recognition"	0

Big Tech for low-resource languages

- **chatGPT** same question in English and Amharic
- The answer to the Amharic question is **totally rubbish, meaningless!**

SE

List all Ethiopian leaders in order



1. Menelik II - Emperor from 1889-1913
2. Lij Iyasu - Emperor from 1913-1916
3. Zewditu - Empress from 1916-1930
4. Haile Selassie I - Emperor from 1930-1974
5. Mengistu Haile Mariam - President from 1974-1987 and Chairman of the Derg from 1977-1991
6. Tesfaye Gebre Kidan - Chairman of the Military Administrative Council from 1987-1991
7. Meles Zenawi - Prime Minister from 1995-2012
8. Hailemariam Desalegn - Prime Minister from 2012-2018
9. Abiy Ahmed - Prime Minister from 2018-present



SE

የኢትዮጵያን መሪዎች በቅደም ተከተል ዘርዝር



አባ ማርያም, ድንቅ ነገሥት አርአይጥ ገብረ አግዚአብሔር, ምንጭ አብዲአስ, ስምዖን አሜሪካ, ሃያላንድ ሰላም, የድሮ ዳርቻ አብዲስ አልማዝ, ሞላዊ ገዳማት መንግሥት ምስራቅ, የውሃ ግንባታ አስተዳደር አህመድ, የውድ ግርማ ወይም የአውሮፓ ነገሥ ዶክተር ጊዮርጊስ መካከለኛ ነገሮች ናቸው።



Recent initiatives - EthioNLP

- Established in 2018 (COLING 2018, Santa Fe, USA)
- Around **40 members** as of today
- It was not active for the last 4 years, getting more active now
- GitHub - <https://ethionlp.github.io/>
- Twitter - <https://twitter.com/EthioNLP>
- Facebook - <https://www.facebook.com/groups/1275577692604176/about/>
- Slack - <https://ethionlp.slack.com/home>
- Telegram - https://t.me/+f_5gMa4KhtU2NwUy



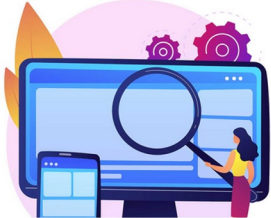
Recent initiatives - GanaNLP




- <https://ghananlp.org/>
- Processing (NLP) of Ghanaian Languages & it's Applications to Local Problems

Our Projects

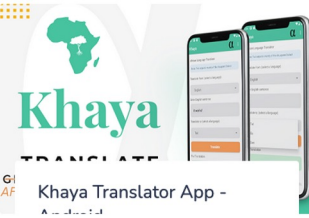
Here are projects contributed to by the community




Khaya Translator Web App



ABENA



Khaya Translator App - Android



Khaya Translator App - IOS

Recent initiatives - HausaNLP



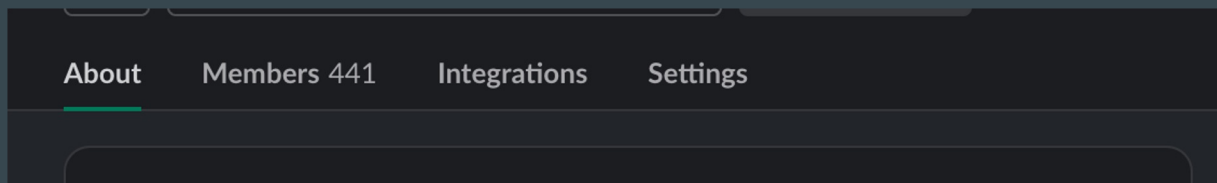
- **Papers**
- **Datasets**
- **Models**
- **Repositories**

- <https://github.com/hausanlp/Awesome-HausaNLP>
- Collaborate with EthioNLP for AfriHate and AfriSenti Projects

Recent initiatives - Maskhane

A grassroots NLP community for Africa, by Africans

- <https://www.masakhane.io/>



Values

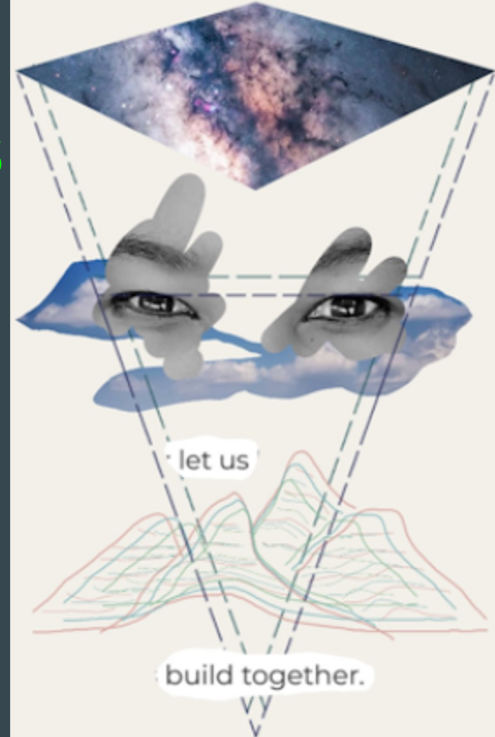
Umuntu Ngumuntu Ngabantu

African-centricity Ownership Openness

Multidisciplinarity Everyone has valuable knowledge

Kindness Responsibility Data sovereignty

Reproducibility Sustainability



Semantic models for Amharic

- Purposes
 - Benchmark Datasets
 - Open sources (models, codes, tools, data)

Announcements

🎉🎉🎉 The Amharic RoBERTa model is uploaded in Huggingface [Amharic RoBERTa Model](#) 🎉🎉🎉

🎉🎉 The Amharic FLAIR embedding model is integrated into the FLAIR library as [am-forward](#) 🎉🎉 The model will be accessible on the next FLAIR release. [Details](#)

🎉🎉 The Amharic Segmenter, Tokenizer, and Transliterator is released and can be installed as `pip install amseg` 🎉🎉

🎉🎉 The Flair based Amharic NER classifier model is now released [am-flair-ner](#) 🎉🎉

🎉🎉 The Flair based Amharic Sentiment classifier model is now released [am-flair-sent](#) 🎉🎉

🎉🎉 The Flair based Amharic POS tagger is now released [am-flair-pos](#) 🎉🎉

Different semantic models and applications for Amharic



AMHARIC NLP
BENCHMARK RESOURCES

<https://github.com/uhh-lt/ethiopicmodels>



Semantic models for Amharic (Yimam et al. 2021)

- Corpus
 - At the [Mendeley Dataset Repository](#)
- Datasets
 - Sentiment analysis
 - NER
 - POS tagging
 - Question classification
- Models
 - Language models
 - AmRoBERTa at Huggingface
 - AmFLAIR - at FLAIR repository
 - Word2Vec
 - fastText
- Segmenter/tokenizer

Hosted inference API ⓘ

Fill-Mask **uhhlt/am-roberta** Example 2 ▾

Mask token: <mask>

የአገሪቱ አጠቃላይ የስነጻዕ አቅርቦት ሶስት አራተኛው የሚመረተው በአገር <mask> ነው።

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.043 s

ወስጥ	0.952
ጸረጽ	0.025
ወስጥ	0.003
ወስጥም	0.003
ወስጥ	0.003

```
pip install amseg
```



EACL 2023



EthioNLP



Natural Language Processing in Ethiopian Languages

Current State, Challenges, and Opportunities

Atnafu Lambebo Tonja , Tadesse Destaw Belay , Israel Abebe Azime , Abinew Ali Ayel,
Moges Ahmed Mehamed, Olga Kolesnikova , Seid Muhie Yimam

<https://github.com/EthioNLP/Ethiopian-Language-Survey>

2.2. POS Tagging

- [Part of Speech tagging for Amharic using Conditional Random Fields](#)
- [Methods for Amharic Part-of-Speech Tagging](#)
- [Amharic Part-of-Speech Tagger for Factored Language Modeling](#)
- [Part of speech tagging for Amharic](#)
- [POS Tagging for Amharic Text: A Machine Learning Approach](#)
- [Part-of-speech tagging for underresourced and morphologically rich languages—the case of Amharic](#)
- [Parts of Speech Tagging for Afaan Oromo](#)
- [Tigrinya Part-of-Speech Tagging with Morphological Patterns and the New Nagaoka Tigrinya Corpus](#)
- [Part of Speech Tagging for Wolaita Language using Transformation Based Learning \(TBL\) Approach](#)
- [A comparative study on different techniques for thai part-of-speech tagging](#)
- [Machine Learning Approaches for Amharic Parts-of-speech Tagging](#)
- [Towards improving Brill's tagger lexical and transformation rule for Afaan Oromo language](#)
- [Deep learning-based part-of-speech tagging of the Tigrinya language](#)
- [Introducing various Semantic Models for Amharic: Experimentation and Evaluation with multiple Tasks and Datasets](#)

2.3. Question Classification and Answering

- [Question Classification in Amharic Question Answering System: Machine Learning Approach](#)
- [Amharic Question Classification System Using Deep Learning Approach](#)
- [Amharic Question Answering for Biography, Definition, and Description Questions](#)
- [TETEYEQ: Amharic Question Answering For Factoid Questions](#)
- [Question Answering Classification for Amharic Social Media Community Based Questions](#)

2.4. Named Entity Recognition (NER)

- [MasakhaNER: Named Entity Recognition for African Languages](#)
- [Amharic Named Entity Recognition Using A Hybrid Approach](#)
- [Named entity recognition for Amharic using deep learning](#)
- [Named Entity Recognition for Amharic Using Stack-Based Deep Learning](#)
- [ANEC: An Amharic Named Entity Corpus and Transformer Based Recognizer](#)
- [Named Entity Recognition for Afaan Oromo](#)

1. NLP Tools

Tools Name	Tools task	Language support	Resource link
amseg	Segmenter, tokenizer, transliteration, romanization and normalization	Amharic	amseg
HornMorpho	Morphological analysis	Amhric, Afaan Ormo, Tigirgna	HornMorpho
lemma	Lemmatizer	Amhric	lemma

2. NLP Applications

2.1. Machine Translation (MT)

We discuss the MT progress for Ethiopian languages in three categories: **English Centeric** -> works done for the above target Ethiopian languages with English pair, **Ethiopian - Ethiopian** -> works done for Ethiopian language pairs without involving other languages and **Multilingual MT** -> works done for Ethiopian languages with other languages in a multilingual setting.

2.1.1 English centeric

- [Parallel Corpora Preparation for English-Amharic Machine Translation](#)
- [Extended Parallel Corpus for Amharic-English Machine Translation](#)
- [Context based machine translation with recurrent neural network for English-Amharic translation](#)
- [Offline Corpus Augmentation for English-Amharic Machine Translation](#)
- [The Effect of Normalization for Bi-directional Amharic-English Neural Machine Translation](#)
- [Optimal Alignment for Bi-directional Afaan Oromo-English Statistical Machine Translation](#)
- [English-Afaan Oromo Statistical Machine Translation](#)
- [English-Oromo Machine Translation: An Experiment Using a Statistical Approach](#)
- [Crowdsourcing Parallel Corpus for English-Oromo Neural Machine Translation using Community Engagement Platform](#)
- [Machine Learning Approach to English-Afaan Oromo Text-Text Translation: Using Attention based Neural Machine Translation](#)
- [The effect of shallow segmentation on English-Tigrinya statistical machine translation](#)
- [Morphological Segmentation for English-to-Tigrinya Statistical Machine Translation](#)

Survey - conclusion

- We investigated the progress of common NLP tasks in Ethiopian languages
- We explored the main NLP research directions, progress, challenges, and opportunities for Ethiopian languages.
- Low prevalence of valuable resource publications (the majority are MSc thesis)
- The publicly available datasets, models, and tools are released in a centralized GitHub repository: <https://github.com/EthioNLP/Ethiopian-Language-Survey>

Why **STILL** low-resource, see MT as an example for **Amharic**

2003 application of corpus-based techniques to amharic texts

2006 guarani: a case study in resource development for quick ramp-up mt

2017 amharic-english speech translation in tourism domain

2018 parallel corpora for bi-lingual english-ethiopian languages statistical machine translation

2018 parallel corpora for bi-directional statistical machine translation for seven ethiopian language pairs

2019 english-ethiopian languages statistical machine translation

2019 language modelling with nmt query translation for amharic-arabic cross-language information retrieval

2022 geezswitch: language identification in typologically related low-resourced east african languages

2022 extended parallel corpus for amharic-english machine translation

Application of corpus-based techniques to Amharic texts

Sisay Fissaha and Johann Haller

Institute for Applied Information Sciences– University of Saarland

Martin-Luther-Str.14, D-66111, Saarbrücken, Germany

Tel +49-681-3895126, Fax +49-681-3895140

{sisay, hans}@iai.uni-sb.de

<http://www.iai.uni-sb.de>

No mention of “low-
resource”

2003

Abstract

A number of corpus-based techniques have been used in the development of natural language processing application. One area in which these techniques have extensively been applied is lexical development. The current work is being undertaken in the context of a machine translation project in which lexical development activities constitute a significant portion of the overall task. In the first part, we applied corpus-based techniques to the extraction of collocations from Amharic text corpus. Analysis of the output reveals important collocations that can usefully be incorporated in the lexicon. This is especially true for the extraction of idiomatic expressions. The patterns of idiom formation which are observed in a small manually collected data enabled extraction of large set of idioms which otherwise may be difficult or impossible to recognize. Furthermore, preliminary results of other corpus-based techniques, that is, clustering and classification, that are currently being under investigation are presented. The results show that clustering performed no better than the frequency base line whereas classification showed a clear performance improvement over the frequency base line. This in turn suggests the need to carry out further experiments using large sets of data and more contextual information.

In this paper, it is mentioned 4X “low-resource”

Extended Parallel Corpus for Amharic-English Machine Translation

Andargachew Mekonnen Gezmu, Andreas Nürnberger, Tesfaye Bayu Bati

Abstract

2022

This paper describes the acquisition, preprocessing, segmentation, and alignment of an Amharic-English parallel corpus. It will be helpful for machine translation of a low-resource language, Amharic. We freely released the corpus for research purposes. Furthermore, we developed baseline statistical and neural machine translation systems; we trained statistical and neural machine translation models using the corpus. In the experiments, we also used a large monolingual corpus for the language model of statistical machine translation and back-translation of neural machine translation. In the automatic evaluation, neural machine translation models outperform statistical machine translation models by approximately six to seven Bilingual Evaluation Understudy (BLEU) points. Besides, among the neural machine translation models, the subword models outperform the word-based models by three to four BLEU points. Moreover, two other relevant automatic evaluation metrics, Translation Edit Rate on Character Level and Better Evaluation as Ranking, reflect corresponding differences among the trained models.

Why we are have more “low-resource” terms over time

- “Low-resource” - becomes **buzzword**
 - Funding
 - Research gap, an opportunity for students
- English and other languages are getting more attention
- The works are less impactful
 - Unpublished
 - Not used in industry

Publish and Perish



Sentiment Analysis



Negative

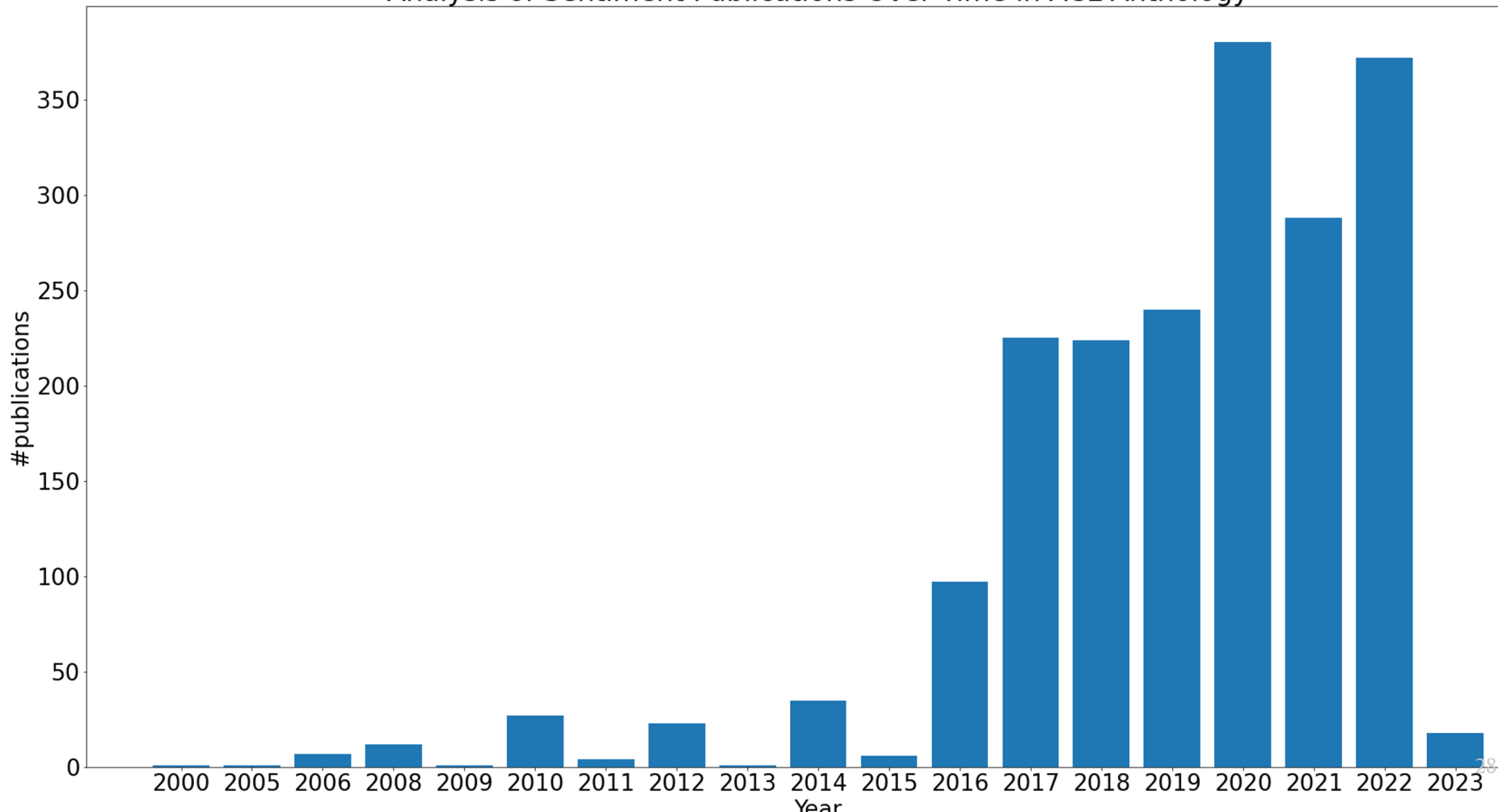


Neutral



Positive

Analysis of Sentiment Publications Over Time in ACL Anthology



Why sentiment analysis is important

- No exception: understand users opinion towards a target
- But, why focus on low-resource languages
 - Difficult to get suggestions/recommendations from multiples sources - **low-resource**
 - Opinions are culturally different - communities have their own language to understand a text
 - Understand opinions for local events, **disaster**, conflict,

ASAB - Amharic Sentiment Analysis (Yimam et al. 2020)

- Sentiment analysis dataset for **Amharic**
- Using **AmTweet dataset**
- Annotation **tools, models, and datasets**

tweet id	tweet	sentiment
120040204070000	የታዘቅ	Positive
120140204070000	የታዘቅ	Positive
120240204070000	የታዘቅ	Positive
120340204070000	የታዘቅ	Positive
120440204070000	የታዘቅ	Positive
120540204070000	የታዘቅ	Positive
120640204070000	የታዘቅ	Positive
120740204070000	የታዘቅ	Positive
120840204070000	የታዘቅ	Positive
120940204070000	የታዘቅ	Positive
121040204070000	የታዘቅ	Positive
121140204070000	የታዘቅ	Positive
121240204070000	የታዘቅ	Positive
121340204070000	የታዘቅ	Positive
121440204070000	የታዘቅ	Positive
121540204070000	የታዘቅ	Positive
121640204070000	የታዘቅ	Positive
121740204070000	የታዘቅ	Positive
121840204070000	የታዘቅ	Positive
121940204070000	የታዘቅ	Positive
122040204070000	የታዘቅ	Positive

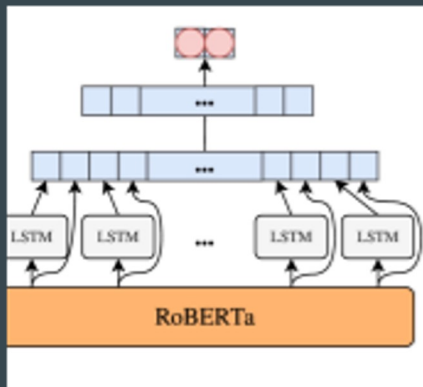
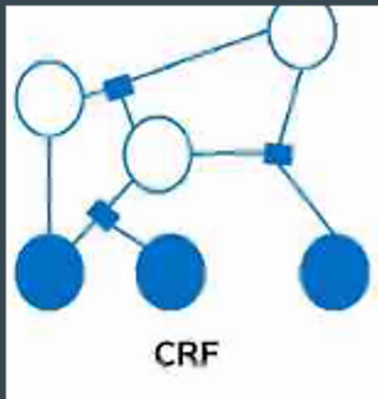
(a) Excel-sheet for annotation

(b) Web annotator interface

(c) ASAB interface

ASAB approach

Filter tweets that are written in Fidel (ⲘⲚⲗ) script



Annotate using ASAB (three users)



Building supervised and deep learning ML models

ASAB tool - <https://github.com/uhh-lt/ASAB>

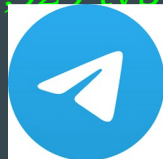
- ASAB support **mobile card vouchers** rewards for annotators.
- Reward given when a user **annotates 50 tweets**.
- ASAB integrates a **controlling control questions** for every 6 tweets.
- Users with 3 consecutive mistakes will receive a **warning** message.
- Users **blocked** after the fourth wrong attempt.

ASAB ML models

- Baseline methods:
 - Stratified, Uniform, and Most frequent.
- Supervised approaches:
 - SVM, KNN, Logistic regression, Nearest centroid
 - Features: TF-IDF with the CountVectorizer and TfidfTransformer methods from scikit-learn.
- Deep learning approaches:
 - Models based on FLAIR deep learning text classifier.
 - Features: Word2Vec, network embeddings, contextual embeddings (RoBERTa and FLAIR embeddings)

Outcome

- 9.4k tweets annotated (143,848 words and 45,525 types), each tweet three annotators.
- A total of 92 Telegram users visited ASAB.
- 58% of users completed at least 50 tweets and got rewarded.
- 4 users blocked for consecutive mistakes.

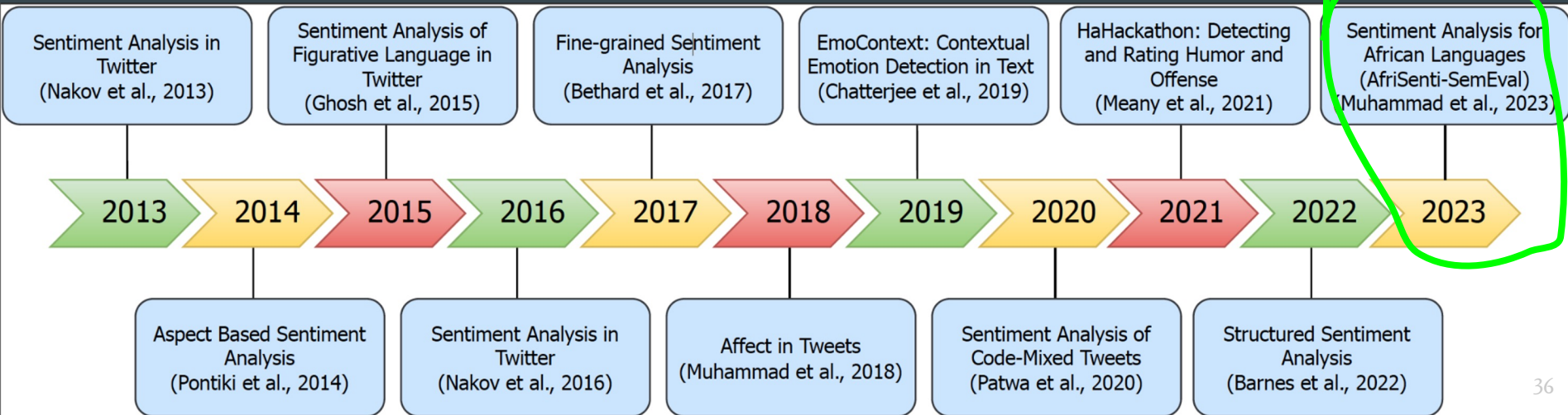


Analysis

- We randomly select tweets where the **model prediction** and the **user annotations** differ.
- Possible source of errors:
 - Users press the **wrong button** by mistake.
 - Some users might not **understand the tweet**.
 - **Slow internet connection**, some users reported that there was a delay between the first and the second tweet.
 - **Sarcasm, figurative speech, mixed scripts**, incomplete phrases and sentences, and spelling and grammar errors cause most of the model errors.

ASAB dataset - extension

- Used for **AfriSenti-SemEval Shared Task 12 - 2023**
- Data is used for the **Amharic Semantic model** project (Yimam et al. 2021)
- ASAB tool is being extended for **general-purpose text annotation**



ASAB model - example usage

Model is currently hosted at the **LT Group** data server

```
import wget
import flair
from flair.data import Sentence
am_sent_model = wget.download("http://ltdata1.informatik.uni-hamburg.de/amharic/taskmodels/sent/final-model.pt")
```

39% [.....]

] 197582848 / 503849408

```
# create example sentence
sentence = Sentence('የብርሃኑ ምርጫ ለኢትዮጵያውያን አሜሪካውያን ስለአገራቸው በደል በቁጣ የሚናገሩበት ይሆናል!')
```

```
# predict class and print
from flair.models import TextClassifier
classifier = TextClassifier.load(am_sent_model)
classifier.predict(sentence)
print(sentence.labels)
```

```
['Sentence[10]: "የብርሃኑ ምርጫ ለኢትዮጵያውያን አሜሪካውያን ስለአገራቸው በደል በቁጣ የሚናገሩበት ይሆናል!" / 'POSITIVE' (0.8838)']
```

AfriSenti-SemEval Shared Task 12 - 2023

AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages

**Shamsuddeen Hassan Muhammad^{1,2+}, Idris Abdulmumin³⁺, Abinew Ali Ayele⁴,
Nedjma Ousidhoum⁵, David Ifeoluwa Adelani^{6*}, Seid Muhie Yimam⁷, Ibrahim Sa'id Ahmad²⁺,
Meriem Beloucif⁸, Saif M. Mohammad⁹, Sebastian Ruder¹⁰, Oumaima Hourrane¹¹, Pavel Brazdil¹,
Felermino Dário Mário António Ali¹, Davis David¹², Salomey Osei¹³, Bello Shehu Bello²,
Falalu Ibrahim¹⁴, Tajuddeen Gwadabe^{*+}, Samuel Rutunda¹⁵, Tadesse Belay¹⁶,
Wendimu Baye Messelle⁴, Hailu Beshada Balcha¹⁷, Sisay Adugna Chala¹⁸,
Hagos Tesfahun Gebremichael⁴, Bernard Opoku¹⁹, Steven Arthur¹⁹**

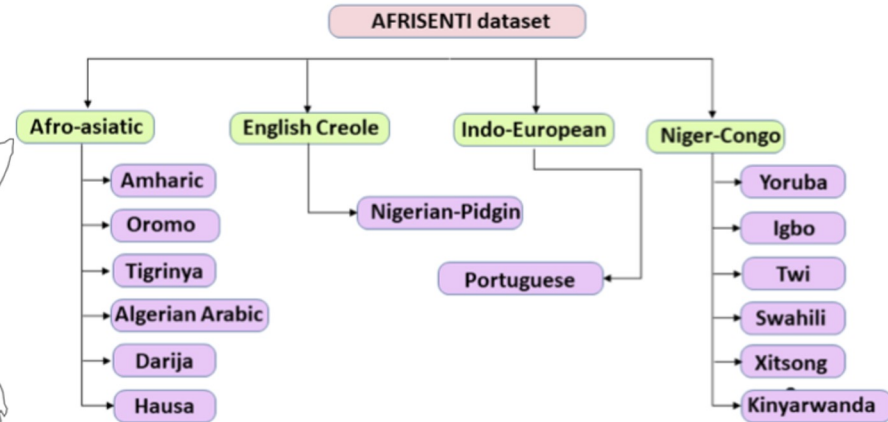
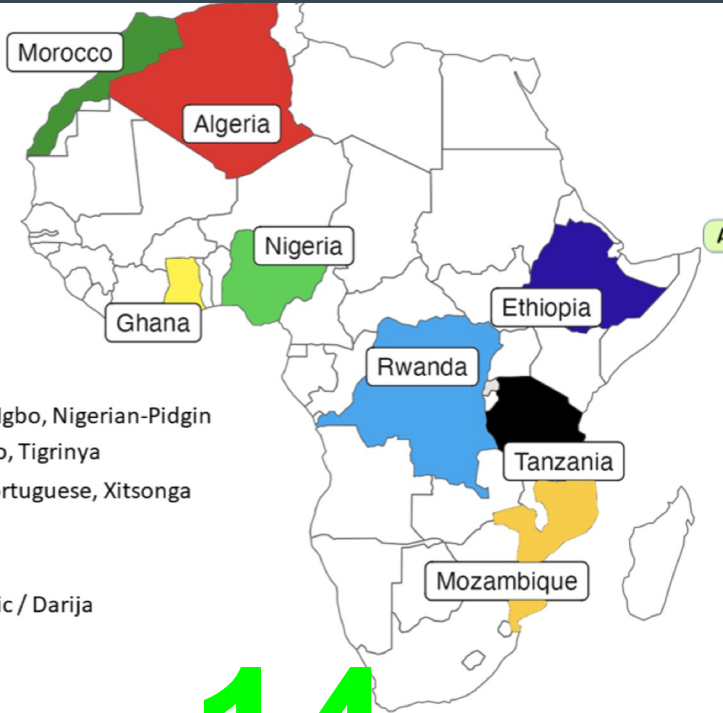
¹University of Porto, ²Bayero University Kano, ³Ahmadu Bello University, Zaria, ⁴Bahir Dar University, ⁵University of Cambridge,

⁶University College London, ⁷Universität Hamburg, ⁸Uppsala University, ⁹National Research Council Canada, ¹⁰Google Research,

¹¹Hassan II University of Casablanca, ¹²dLab, ¹³University of Deusto, ¹⁴Kaduna State University, ¹⁵Digital Umuganda,

¹⁶Wollo University, ¹⁷Jimma University, ¹⁸Fraunhofer FIT, ¹⁹Accra Institute of Technology, *Masakhane NLP, +HausaNLP

AfriSenti datasets



14

Languages

110,000

Afrisenti Dataset collection challenges

- Lack of support for certain African languages letters on keyboards
 - E.g Twi : ε, ɔ
- Code-mixing
 - Between low-resource languages. E.g Yoruba and Igbo
- Tonality challenge
 - Àwon omó fò abó (The children washed the dishes) has a positive meaning,
 - Àwon omó fọ abó (The children broke the dishes) has a negative meaning

Winning system

NLNDE

No Language No Data Expertise

Bosch Center for Artificial Intelligence,
Renningen, Germany; Center for
Information and Language Processing
(CIS), LMU Munich, Germany

The LAPT approach involved **continue pre-training** a PLM on the **monolingual corpus** of a target African language,

TAPT involved **continue pre-training** on the training dataset of the task i.e. **AfriSenti training corpus** of a target language

guage. By leveraging LAPT followed by TAPT, they achieved significant improvements over fine-tuning AfroXLMR-large directly. **NLNDE ranked first in 7 out of 12 languages, and first in sub-task A.**

Hate Speech





Addressing hate speech on social media: Contemporary challenges

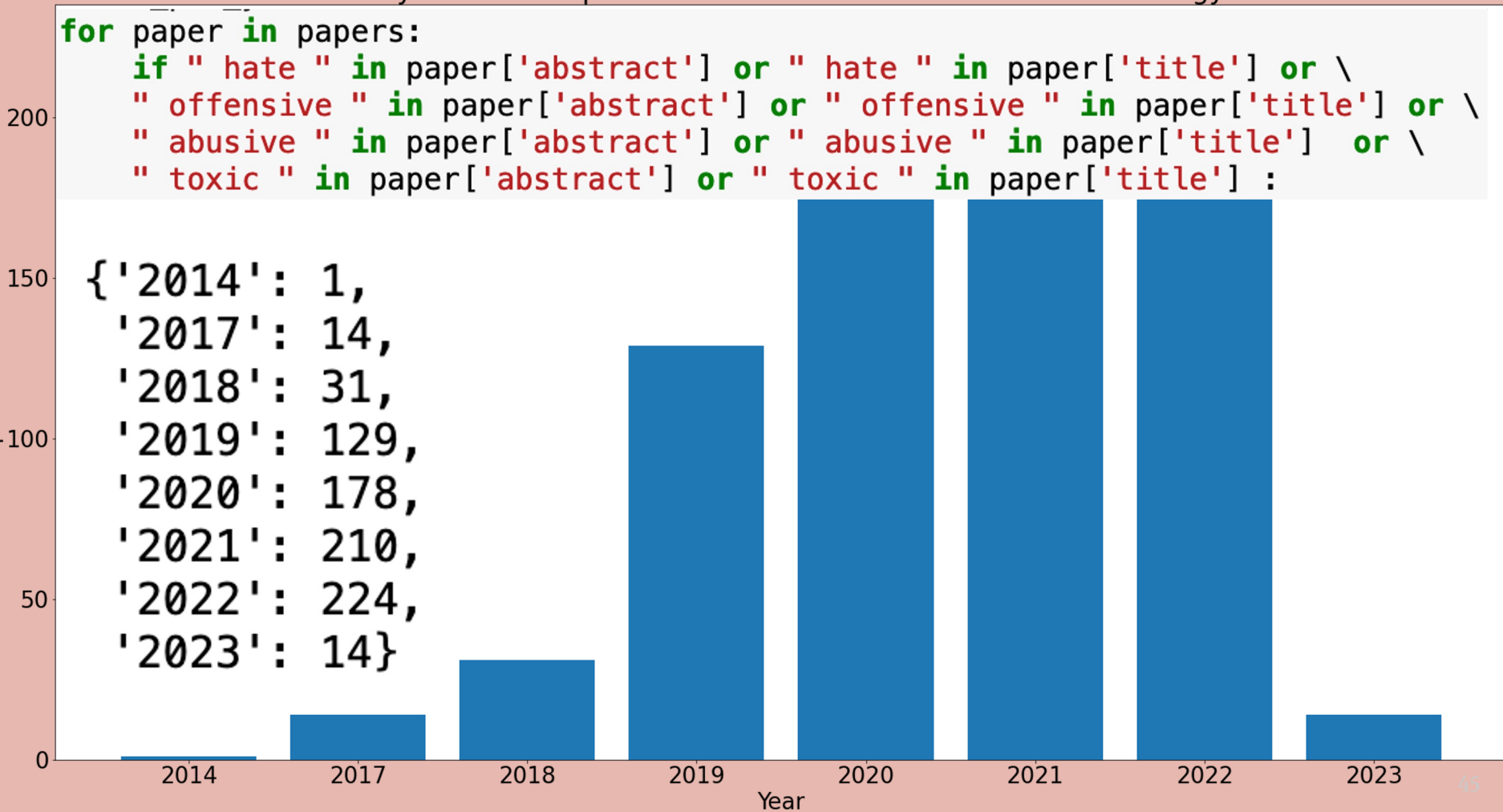
- Freedom of speech?
- “Six part threshold test” to justify restrictions on freedom of expression
 - considering the **socio political** context,
 - status of the **speaker**,
 - intent to incite **antagonism**,
 - speech **content**,
 - extent of **dissemination**
 - likelihood of **harm**.

Analysis of Hate Speech Publications Over Time in ACL Anthology

```
for paper in papers:  
    if " hate " in paper['abstract'] or " hate " in paper['title'] or \  
    " offensive " in paper['abstract'] or " offensive " in paper['title'] or \  
    " abusive " in paper['abstract'] or " abusive " in paper['title'] or \  
    " toxic " in paper['abstract'] or " toxic " in paper['title'] :
```

```
{'2014': 1,  
'2017': 14,  
'2018': 31,  
'2019': 129,  
'2020': 178,  
'2021': 210,  
'2022': 224,  
'2023': 14}
```

#publications



Is **hate speech** a new event?

Ask ChatGPT! :-)



No, hate speech is not a new event. Hate speech has existed for centuries and has been used to target and discriminate against individuals or groups based on their race, religion, ethnicity, gender, sexual orientation, and other characteristics. The term "hate speech" itself may be relatively new, but the concept and practice of using language to express hatred or prejudice towards others is not new. With the rise of social media and other digital platforms, however, hate speech has become more visible and widespread in recent years.



Why **hate speech** detection models for **low-resource** languages?

- **Social media** and mobile usage increases
 - Less awareness of media usage
 - Social media becomes a **fertile ground** for hate speech
 - Escalate to **offline unrest** (Nkemelu et al. 2023)
- Utilize knowledge of experts of the **community** along models
- Plan **counter-hate speech** - peace build up - responsible AI
- But
 - **Small** and **unbalanced** dataset
 - **Unpublished** resources/models

Hate speech in Ethiopia - the 5J

- Use the **Toloka Yandex** crowdsourcing
- **Crowdsourcing** is getting more popular for data annotation due to its **lower cost**, **higher speed**, and diversity of opinions



5Js - Unpacking Ethiopia's Controversial Five Consecutive Junes: A Period of Turmoil and Change (Ayele et al. 2022)

The 5
consecutive
&
controversial
Ethiopian
Junes: 5Js

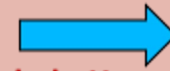
- **June 2018**
Bomb attack at a rally



- **June 2019**
Assassination of Officials



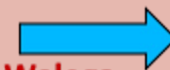
- **June 2020**
Assassination of Hachalu H.



- **June 2021**
National Election 2021



- **June 2022**
Massacre in Kelem Wolega



Data collection and annotation

- Toloka setups:
 - 20 training tweets
 - 50 control tweets
 - Smart mixing:
 - a task has 15 tweets
 - The one is a control question



COLLECTION SPAN

- ↔ 5 Controversial June events, in 5 years
- ↔ approximately 1k tweets for each June



Toloka



ANNOTATION

- ↔ 5.4k tweets
- ↔ 2 pilots & 5 pools
- ↔ Each tweet labelled by 3 annotators
- ↔ Gold label: majority voting

Fleiss Kappa Agreement



- **Pilot1: 0.15,**
- **Pilot2: 0.25 and**
- **Main Task: 0.34**

Sample Toloka user interface for presenters for performers

The screenshot displays the Toloka user interface for performers. At the top, there are navigation tabs for 'Tasks', 'Active', and 'Messages'. The top right corner shows the current time '18:36', earnings '\$0.10' (with 'የጥላቻ ንግግር' below it), a balance of '\$0.00 / \$0.00', and an 'Instructions' link. The main area contains a grid of task cards. Each card features a quote in Amharic, a classification question 'How would you classify the tweet?', and radio button options for 'hate', 'offensive', 'normal', and 'unsure'. Some cards also include a question 'Against whom is the hate or offense directed?' with options for 'racial target', 'non-racial target', and 'unsure'. The interface is clean and organized, with a light green header for each task card.

Tasks Active Messages

18:36 / \$0.10
የጥላቻ ንግግር

? \$0.00 / \$0.00 Instructions

“ @USER ጥቅሙ የተነካበት፣ ፊሪ፣ ሸንታም፣ ደደብ፣ የጠባብ ሀዋት አሸከር.ሌላ ማን ሊሆን ይችላል!

How would you classify the tweet?

1 hate 2 offensive 3 normal
4 unsure

Against whom is the hate or offense directed?

q racial target w non-racial target e unsure

“ ከብዛህኛው ሰው በሚሰራው ስራው ስጦት አየተደመመና አድናቂው ሁኖ ሳለ በጥላቻና ስብሄሩ ብቻ ሚታገል ዘረኛ አንደሆነ ነው ሚመለከቱት ቢሆንም አሱ ሁሌ በሚሰራው ስራ አድማሚ ነው። አሏህ

How would you classify the tweet?

1 hate 2 offensive 3 normal
4 unsure

“ Cካሽ ተግባር . #meskelsquare #devil በከበረው የሰው ህይወት ሞት እና ደም መፍሰስ ደስታን ለማግኘት የሚደረግ አረመኔያዊ የሰይጣን ተግባር Thank You

How would you classify the tweet?

1 hate 2 offensive 3 normal
4 unsure

“ አንዱት ሰው በሰው ላይ ሰይጣን ይሆናል!!! RIP !

How would you classify the tweet?

1 hate 2 offensive 3 normal
4 unsure

“ ሰበር ወረቀት፣ በዛሬው ዕለት በመስቀል አደባባይ በተደረገው የድጋፍ ሰልፍ ላይ በምብ ተወርውሮ የሞትና የመቁሰል አደጋ ደርሷል። የጉዳቱን መጠን ለማወቅ አልተቻለም። ጉዳዩን አስመልክቶ

How would you classify the tweet?

1 hate 2 offensive 3 normal
4 unsure

“ @USER አንተ የቀን ጅብ! ነግረናቸው ነበር።

How would you classify the tweet?

1 hate 2 offensive 3 normal
4 unsure

Against whom is the hate or offense directed?

q racial target w non-racial target e unsure

“ ለጠ.ሚ አብይ አህመድ በመስቀል አደባባይ የተካሄደው የድጋፍ ሰልፍ አንዲሁም በገምብ ፍንዳታ የደረሰው አደጋ በከፊል ።

“ ሸቦሲያ አፈንዳ አይባል ከሸቦሲያ ጋር ታርቀናል፣ በግንቦት 7 አይሰብብ ግንቦት ሰባት ተሰማምቷል? ኮነግ ጋር ማ ታርቀናል። ታዲያ በንቡን ማን አፈነዳው? የቀን ጅብቹ?

“ ተጨማሪ መረጃ በመስቀል አደባባይ በነበረው ሰልፍ የበምብ ጥቀት ፈፀመዋል ተብለው የተጠረጠሩ 3 ሰዎች መያዛቸውን የአይን ለማኞቻ ተናገሩ። ሰዎቹ የተያዙት

Annotation errors

- Possible source of **variations among human annotators** might be due to:
 - Negligent or malicious annotators working only for financial rewards.



- Tweets containing **idiomatic** and **poetic expression** are difficult to understand
- The **context** in which some tweets are written is not known

Error analysis

#	Tweet	Anno1	Anno2	Anno3	Gold
1	@USER አንተ ደደብ ቁራ...ትህን ግራ። (@USER You idiot. educate your cattle called K...)	<u>normal</u>	<u>normal</u>	offensive	<u>normal</u>
2	...ኛ ከአሮምያ ከኢትዮጵያ ካልጠፋ ሰላም የለም። (If the mu... does not disappear from Oromia and Ethiopia, there will be no peace.)	<u>normal</u>	<u>normal</u>	<u>normal</u>	<u>normal</u>
3	አማራነትን መርጦ የዘር ጭፍጨፋ ማድረግ ይቁም!! (Stop genocide of ethnic Amhara's!!)	<u>hate</u>	<u>hate</u>	<u>hate</u>	<u>hate</u>
4	@USER ተጠያቂነት ካልሰፈነ ጭፍጨፋው ይቀጥላል። (@USER Without accountability, the massacre will continue.)	<u>hate</u>	<u>hate</u>	<u>hate</u>	<u>hate</u>
5	<u>የተበተኑት አውሎ ነፋስ ሆኖ መጣ።</u> (The dis... your comes as a whirlwind.)	<u>normal</u>	<u>normal</u>	unsure	<u>normal</u>
6	@USER አንተ ቀልድ፡ አህያውን ፈርቶ ዳውለውን (@USER you are joking; while fearing the donkey, you deal with what the donkey carries)	<u>hate</u>	<u>hate</u>	<u>hate</u>	<u>hate</u>

Hate

Hate

Normal

Normal

Sarcasm

Idiom

Challenges in hate speech annotation

- **Data selection**: Lexicon? **Party** names? **Ethnics** names? Tricky!!
- **Costly**: no difference from English and similar high-resource languages
- **Sensitive**: Annotators can be annoyed (religion/ethnicity)?
- **Native speaker**: You need speakers who speak the language (annotation, guideline)
- **Awareness**: Training annotators about the implication of the annotation, why do they care?
- **Infrastructure**: Most have mobiles, web-based tools will not help. Where to publish the data (GitHub??)
- **Lack of experts**: There are less researchers in general, and much **worse for NLP**

Lacuna Funding 2022



AfriHate Datasets

Nigeria

Hausa, Igbo, Pidgin, Yoruba

Ethiopia

Amharic, Tigrinya, Oromo, Somali

Algeria

Algerian Arabic

Mozambique

Portuguese

Ghana

Twi, Pidgin

Kenya

Swahili

Sudan

Sudanese Arabic

Rwanda

Kinyarwanda

South Africa

Afrikaans, isiZulu, Isixhosa

Somalia

Somali

Morocco

Darija

Project Leading Universities

Bayero University
Kano, Nigeria



Bahir Dar University,
Ethiopia



Project Partner Organizations



Conclusion

- Most languages, for example Amharic, they are **not anymore low-resource** for some tasks, they are **less-organized**.

"Amharic Machine translation



All



Videos



Images

About 1,340 results (0.55 seconds)

Bahir Dar University
Institutional Repository System

BDU IR Home → Search

Search

Search: All of IR

language processing | Amharic

Add filters

Showing 10 out of a total of 1504 results. (0.026 seconds)

1 2 3 4 ... 151 Next Page

Communities or Collections matching your query

[Ethiopian Languages and Literature - Amharic](#)

Items matching your query



[AUTOMATIC IDIOM RECOGNITION MODEL FOR AMHARIC](#)
ANDUAMLAK, ABEBE FENTA (2021-07)

Language Processing: researche has been influenced by the existence of idioms in natural language. This research shows that idiom affects NLP researches such as machine translation, sentiment analysis, information retrieval, question answering and next word prediction.



[IDIOMATIC EXPRESSION IDENTIFICATION FROM AMHARIC USING DEEP LEARNING](#)
TIRUEDLE, ASTERAYE TSIGE (2022-07)

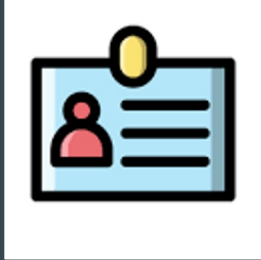
This research proposes a deep learning-based algorithm for identifying idiomatic expressions in Amharic text. The experimental result shows that the proposed algorithm performs better than SVM, KNN, CNN, and LSTM. Keywords: Amharic, Idiomatic Expression, Deep Learning.

58

Conclusion

- Most languages, for example Amharic, they are **not anymore low-resource** for some tasks, they are **less-organized**.
- **Promote** low-resource language works
- Create collaboration among local researchers
- **Mentoring** of students in low-resource language

Question/discussion/contact me?



Seid Muhie Yimam

House of Computing and Data Science

Universität Hamburg

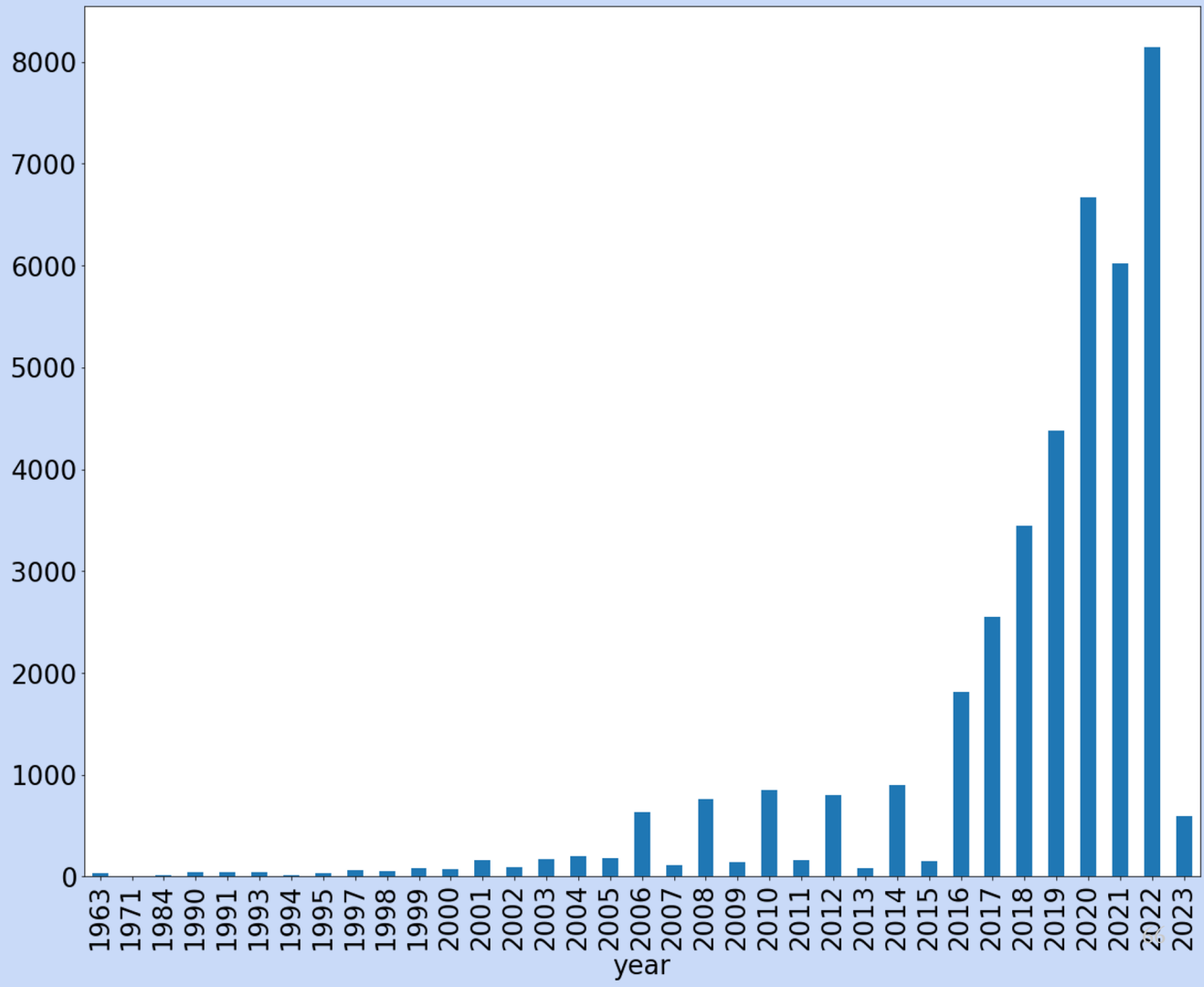


seid.muhie.yimam@uni-hamburg.de



@seyyaw

ACL Anthology papers per year



year	count
2014	85240
2015	346016
2016	433022
2017	498085
2018	695533
2019	1032463

Amharic tweets (AmTweet)- **current status**

- Collect tweets everyday
- Tweets written in Amharic script (**Ethiopic, Fidäl, Ge'ez**)
- A total of **17,602,943 tweets** by April 16, 2023

IPA	æ	u:	i:	a:	e:	ə	o:	wə	jæ
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ		
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ	ሊ	

2020	2291416
2021	3454069
2022	6786826
2023	1979773

