A Sentiment Corpus for South African Under-Resourced Languages in a Multilingual Context

Ronny Mabokela and Tim Schlippe
University of Johannesburg, South Africa
IU International University of Applied Sciences, Germany

24-25 JUNE 2022, SIGUL2022 @LREC2022

Presentation Outline

- Introduction
- Related Work
- Methodology
- SAfriSenti Corpus and Resources
- Corpus Statistics
- Contributions
- Conclusion

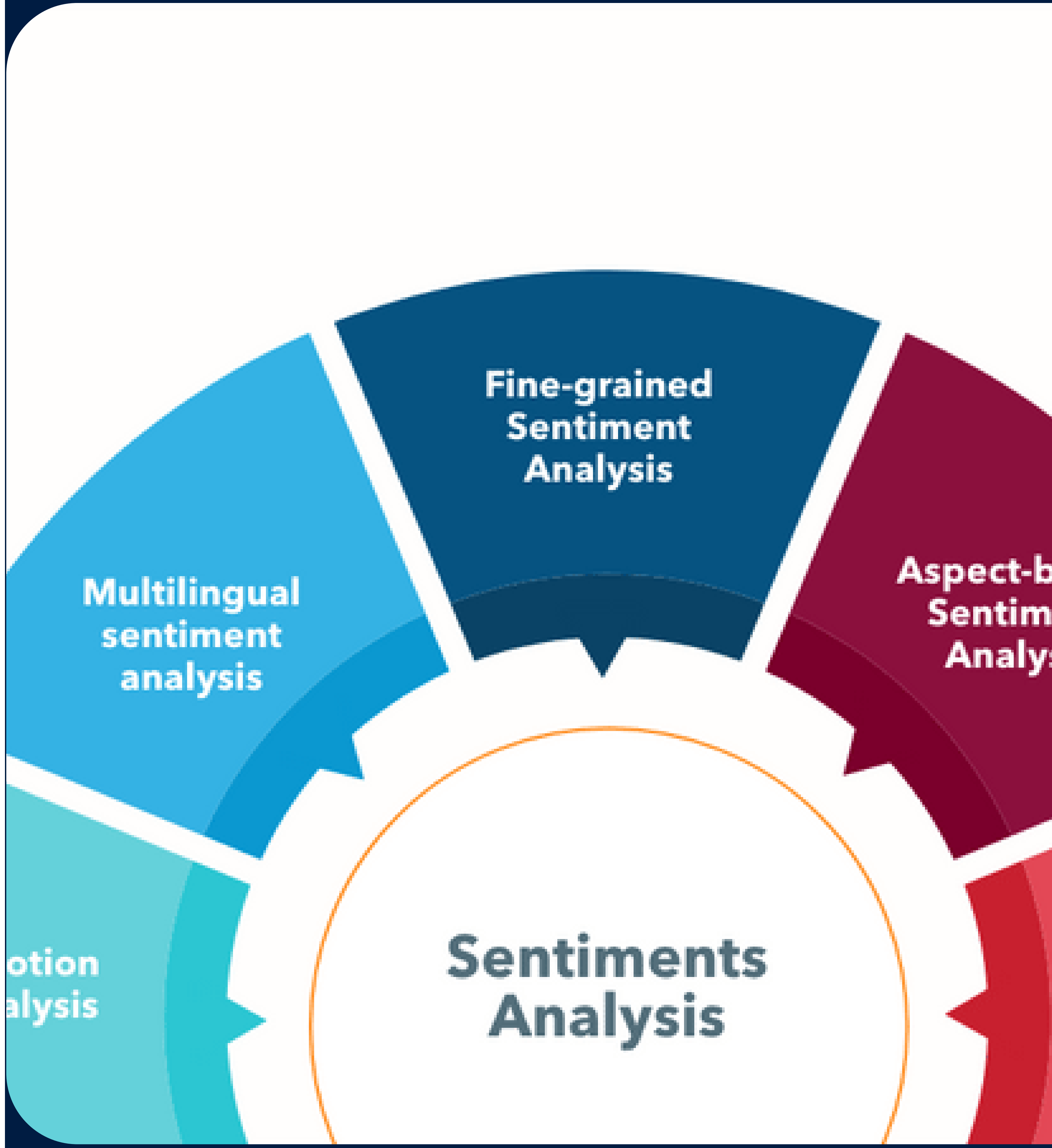
Introduction

Detecting sentiments or emotions from language — A significant area of research in NLP for the past decades.

- **Sentiment analysis (SA) is concerned with detecting and categorising emotions from textual information**

- **SA has earned research attention which may be attributed to numerous essential NLP applications.**

SA has promising progress in high-resource languages, e.g, English and Chinese. But the same cannot be said for languages with limited resources.


Fine-grained Sentiment Analysis
Multilingual sentiment analysis
Aspect-b Sentim Analy
otion alysis
Sentiments Analysis

Introduction...

Lack of resources poses a significant challenge for language-specific services

Under-resourced languages are in desperate need of data, digital tools, and resources

Socio-cultural factors, multicultural factors affect languages

Recently, SA has introduced multilingual sentiment analysis due to the rapid use of a mixture of languages on various social media platforms

Related Work

South African Landscape

- South Africa has over **60 million people**.
- **11 official spoken languages** and over 50 dialects.
- African country with the **sixth-largest population**.
- Most multilingual and multicultural societies:
  - Native speakers are fluent in at least two languages.

Social Media Usage

- A report shows that in 2020 approximately **40%** of South Africa’s population were active on social media platforms and approximately **9.3 million** of those are on **Twitter**.

Related Work...

African Languages

SA for monolingual, code-switched and multilingual comments has been studied for a few African languages:

- Several Nigerian languages
- Swahili
- Bambara

Cross-Lingual SA Methods

- Cross-lingual methods to solve the challenges of under-resourced languages.
- Done by utilising language knowledge from high-resource languages like English.
- Translate the comments from the original language to conduct its classification task with high-performing models that are trained with large English resources

Related Work...

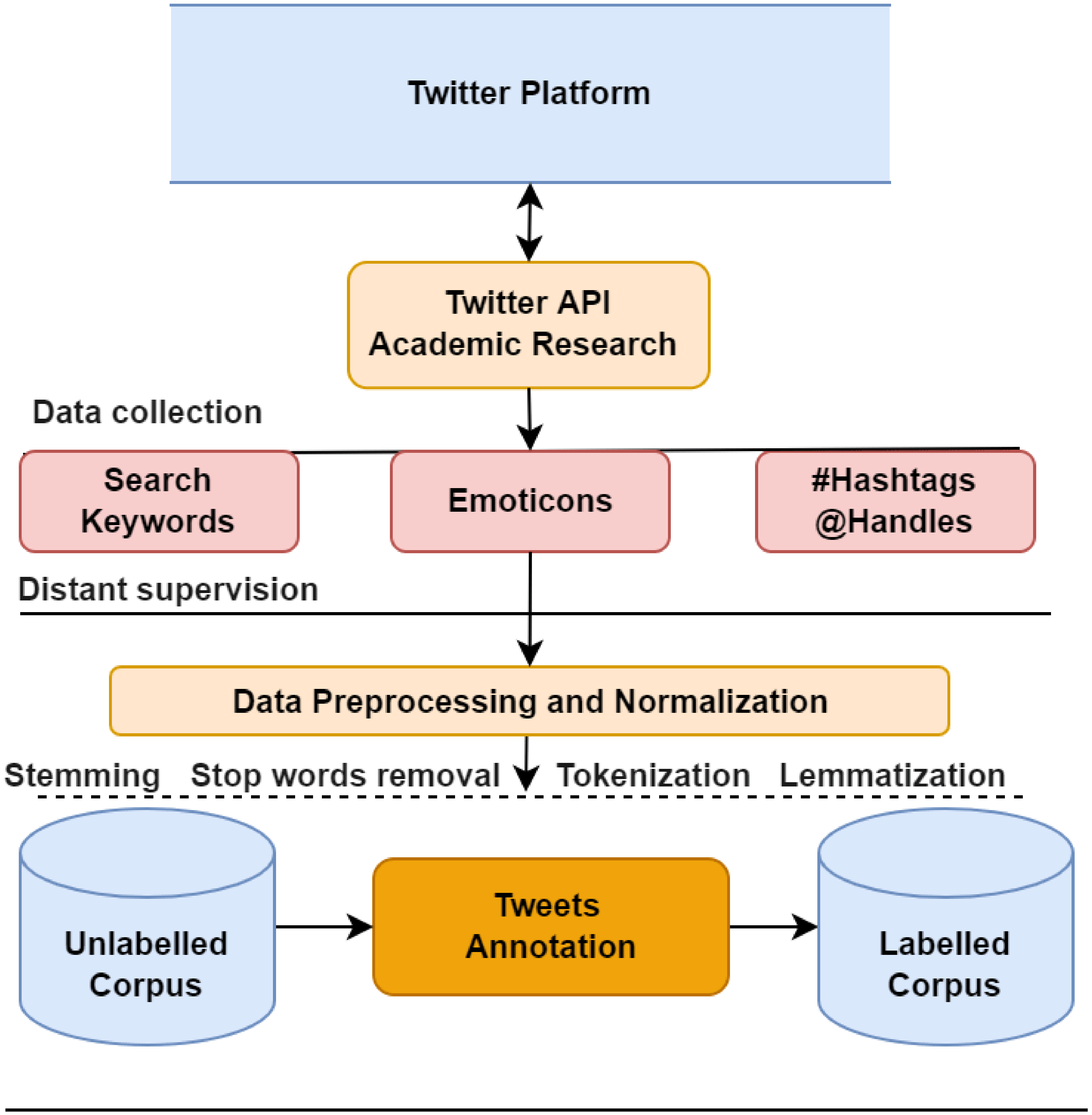
Cross-Lingual SA Methods

- This approach was successful for the high-resourced languages like Chinese, French, Russian, German and Spanish.

- However, translation from English to German, Urdu, and Hindi had a bad impact on SA performance.

- But there was a 2-3% SA performance decrease from English to under-resourced languages with help of MT compared to human translation.

Methodology



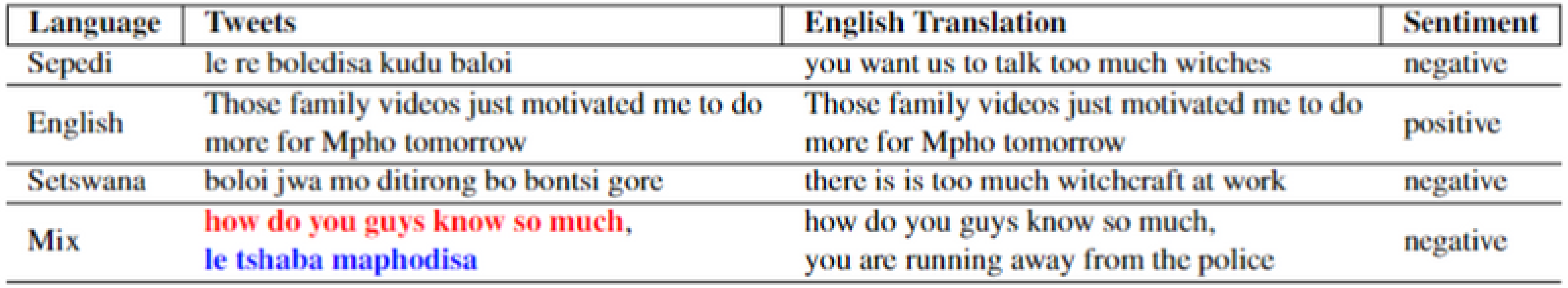
Data Collection

1 **Twitter Data Collection**

2 **Removal of Short and Duplicated Tweets**

3 **Preprocessing and Normalisation**

- Examples: **Loooool** or **Whaaaaat** and **ngwanaaaaaka** is replaced with **Lol** or **What** and **ngwanaka**



Language	Tweets	English Translation	Sentiment
Sepedi	le re boledisa kudu baloi	you want us to talk too much witches	negative
English	Those family videos just motivated me to do more for Mpho tomorrow	Those family videos just motivated me to do more for Mpho tomorrow	positive
Setswana	boloi jwa mo ditirong bo bontsi gore	there is is too much witchcraft at work	negative
Mix	how do you guys know so much, le tshaba maphodisa	how do you guys know so much, you are running away from the police	negative

Table 1: Example of tweets, their corresponding English translation as well as their associated sentiment labels

SAfriSenti Corpus

1 Pre-Annotation

- Emoticons are used as a distantly supervised method to pre-classify tweets as positive, neutral or negative.
- Positive, neutral and negative search keywords.

2 Annotator's background and training

- Recruit annotators: 3 native speakers of each language.
- Technical and linguistic background.
- SentiApp — an online platform for organising and annotating tweets.

3 Annotation Guidelines

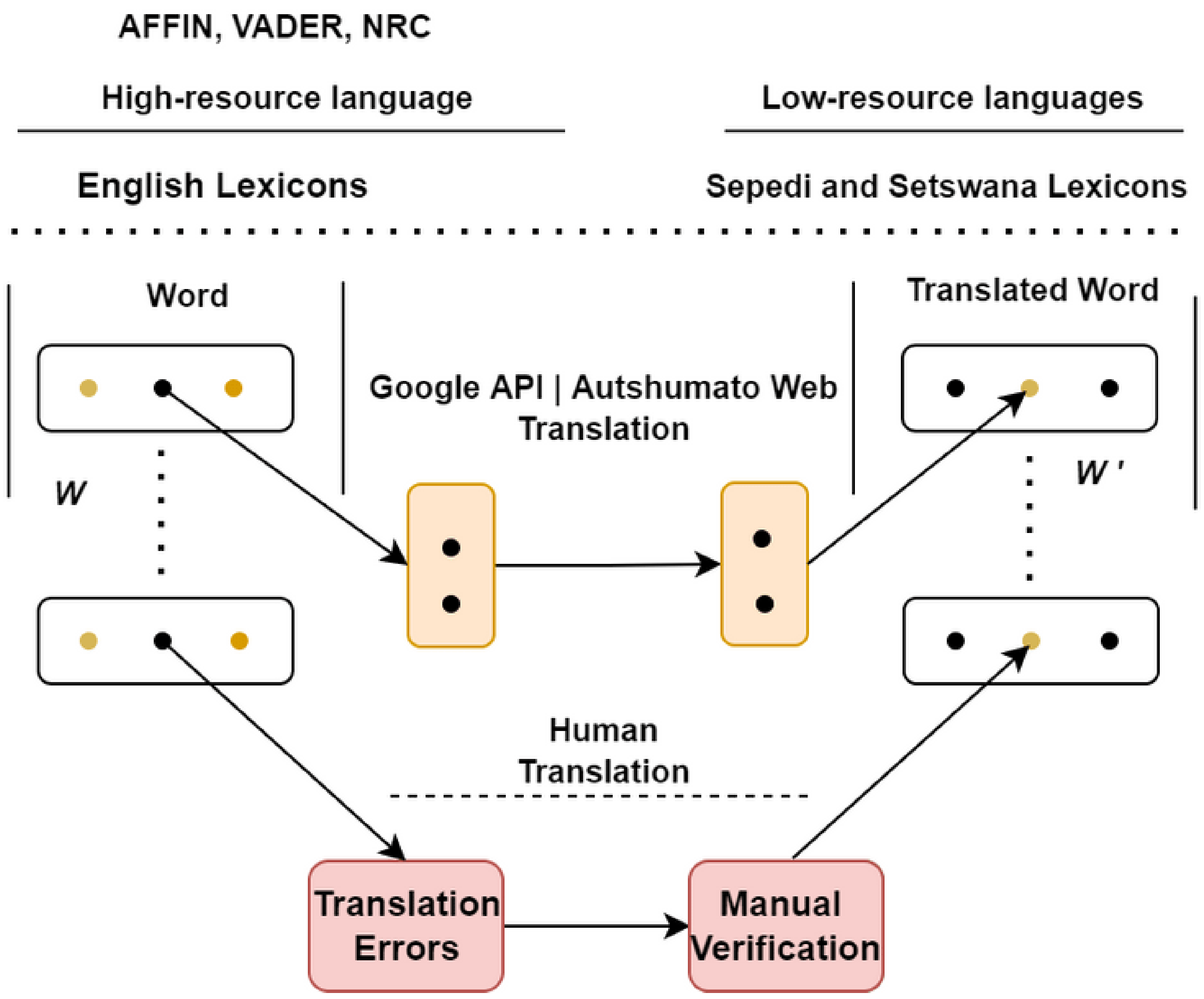
- Positive Sentiment (POS)
- Negative Sentiment (NEG)
- Neutral Sentiment (NEU)
- Positive and Negative Sentiment

4 Annotation Process & Voting

- Three-way disagreement (NEG, NEU and POS).
- Three-way agreement (NEG NEG NEG => NEG).
- Two-way partial disagreement (POS, POS, NEU).
- Two-way disagreement (POS, POS, NEG).

Additional Language Resources

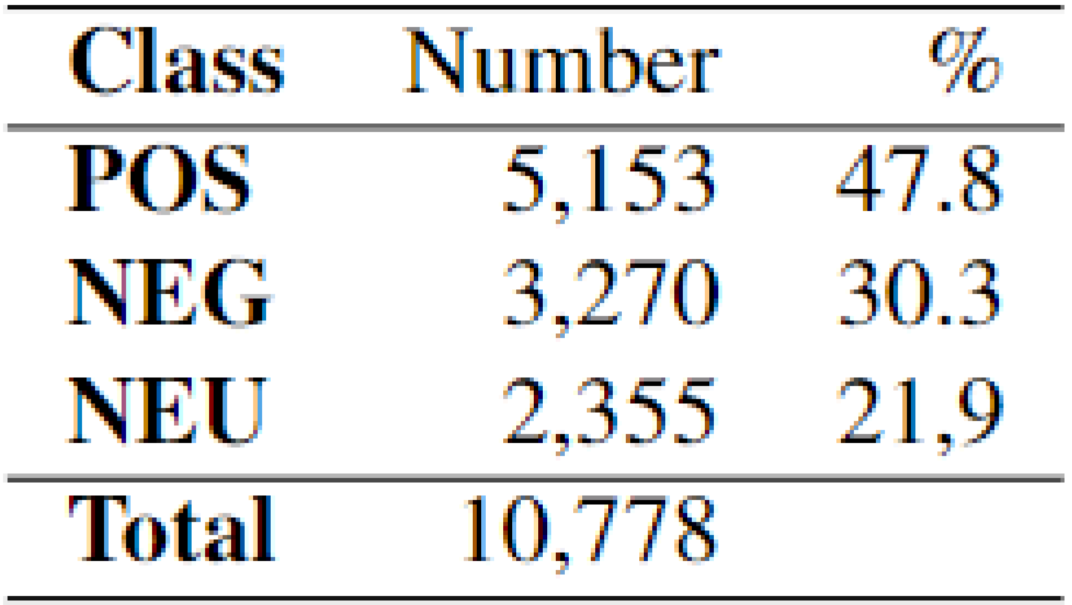
Sentiment Lexicons



Sentiment Taggers

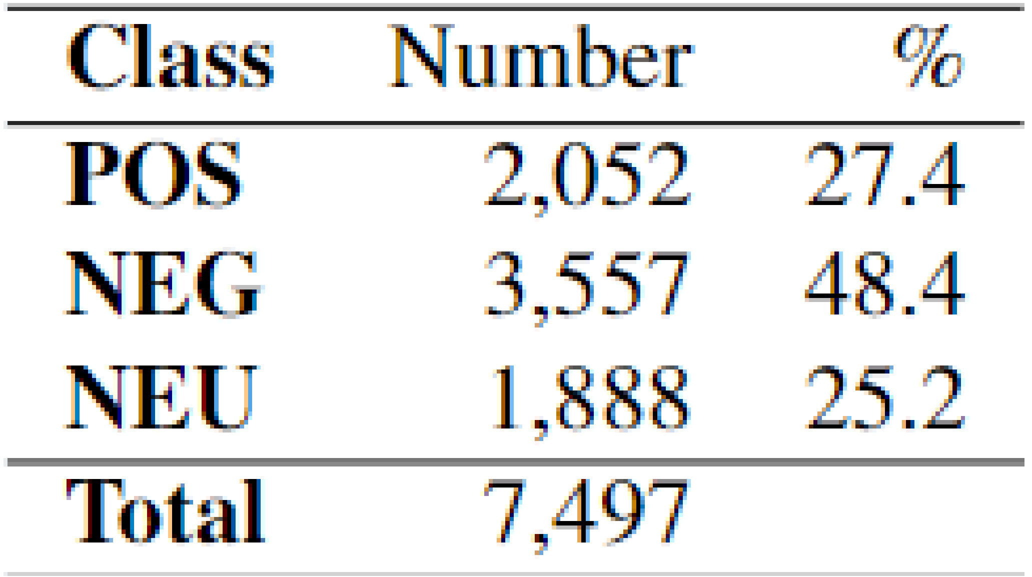
- Sentiment taggers for Sepedi and Setswana
- Examples of Sepedi morphemes which indicate a negative mood are: **/ke be ke sa/** and **/ba be ba sa/**.
- Examples of Sepedi morphemes which indicate a positive mood are: **/ke be ke/** or **/ba be ba/**.

Data Statistics: Monolingual Tweets



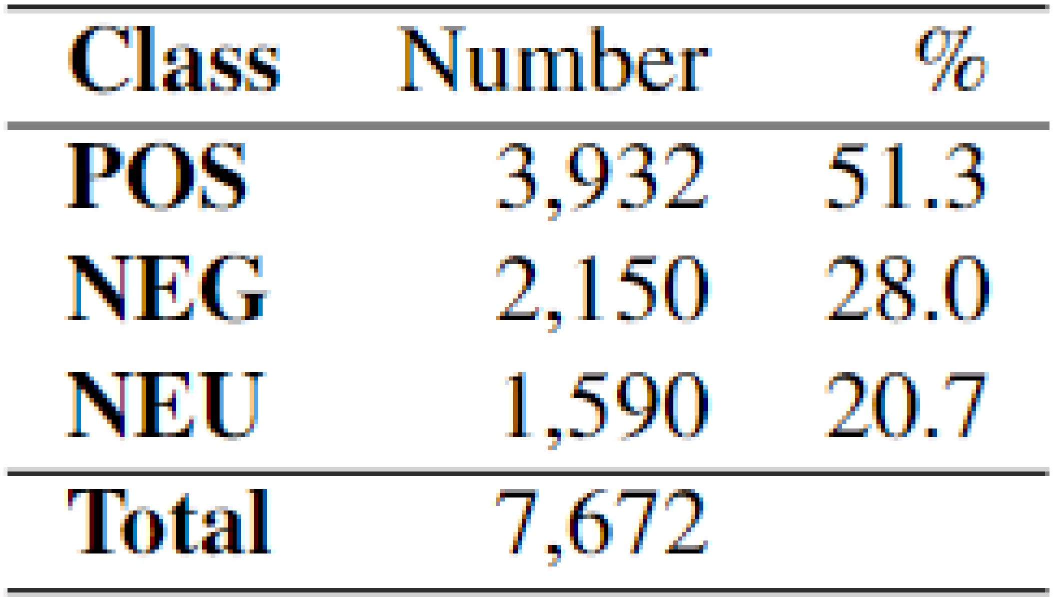
Class	Number	%
POS	5,153	47.8
NEG	3,270	30.3
NEU	2,355	21,9
Total	10,778	

Distribution of Sepedi tweets



Class	Number	%
POS	2,052	27.4
NEG	3,557	48.4
NEU	1,888	25.2
Total	7,497	

Distribution of English tweets

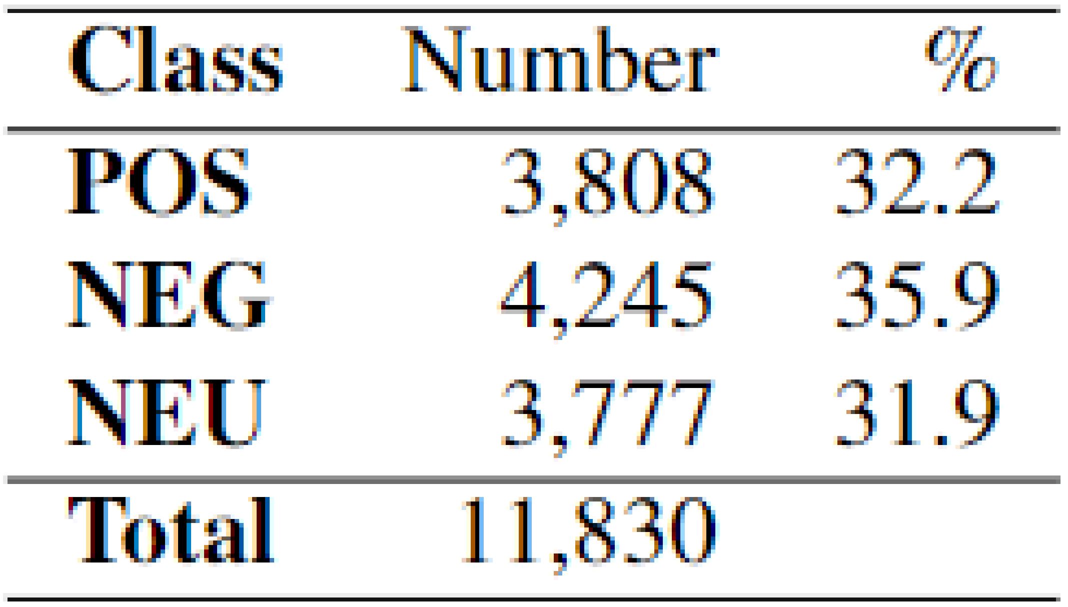


Class	Number	%
POS	3,932	51.3
NEG	2,150	28.0
NEU	1,590	20.7
Total	7,672	

Distribution of Setswana tweets

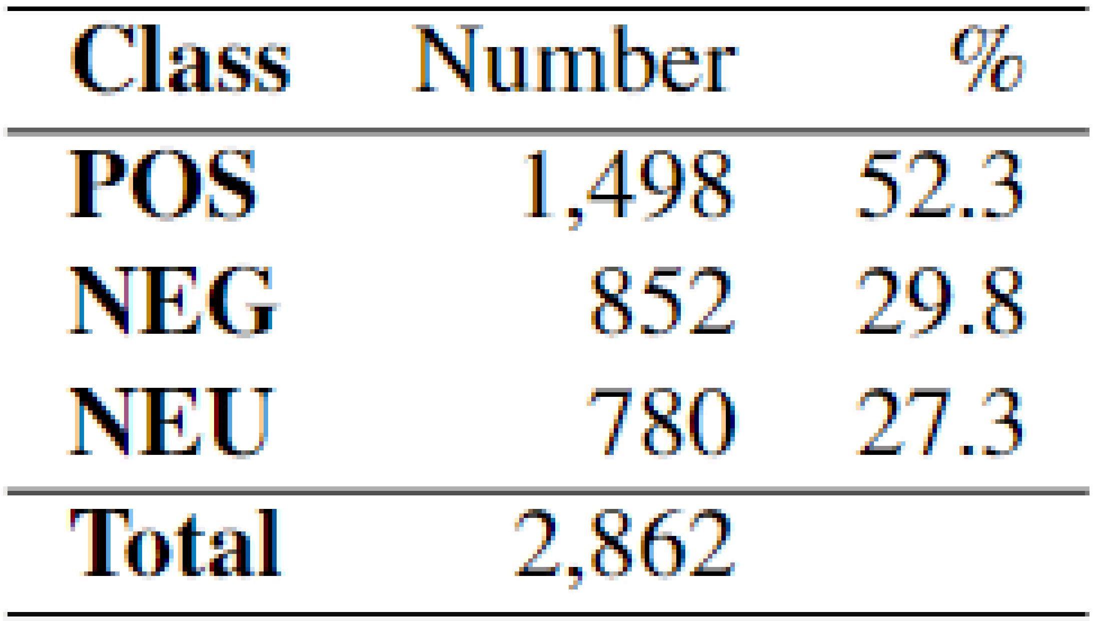
- We report only the annotated subset of over 40,000 tweets.
- The monolingual tweets cover 63.4% (26k tweets).
- Our subset consists of a large number of code-switched tweets (15k tweets).

Data Statistics: Code-Switched Tweets



Class	Number	%
POS	3,808	32.2
NEG	4,245	35.9
NEU	3,777	31.9
Total	11,830	

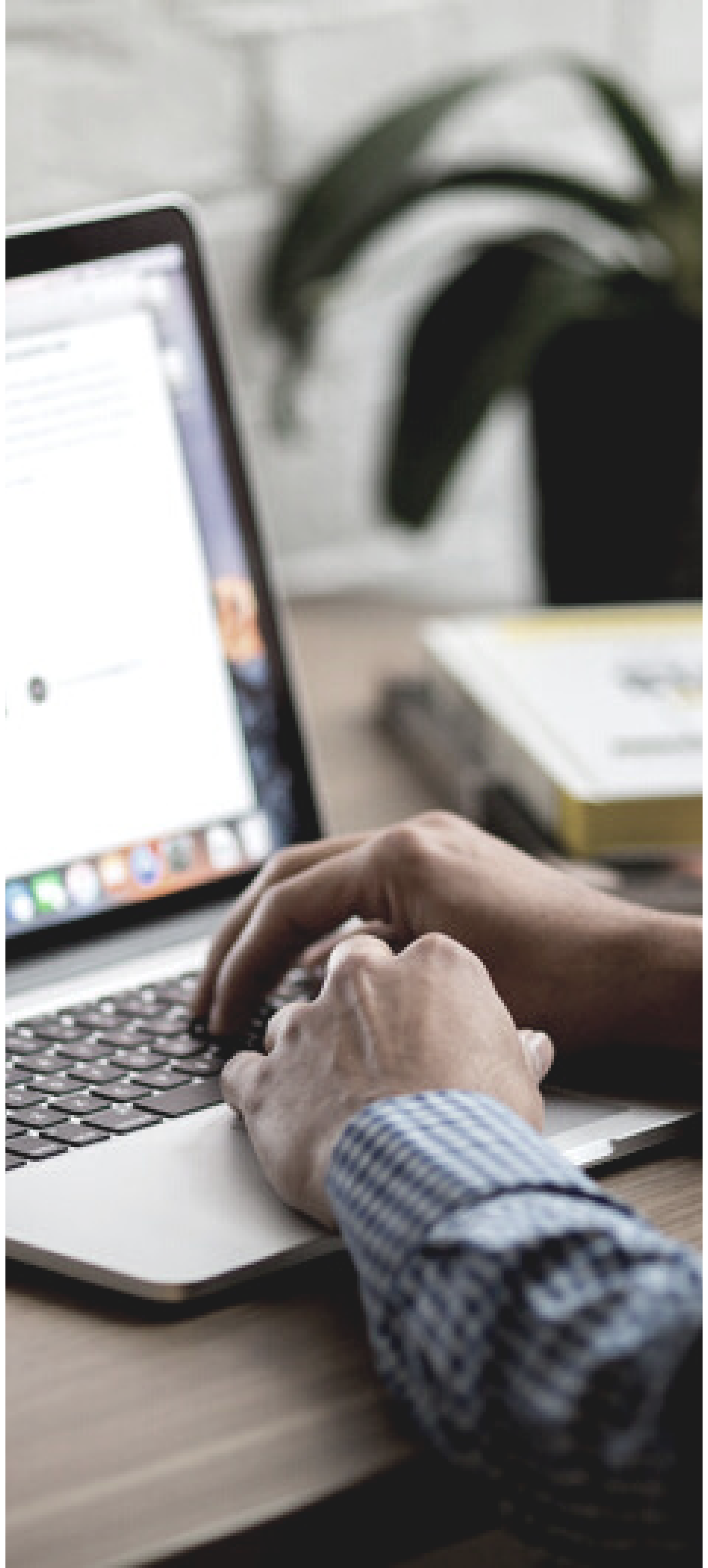
Distribution of English-Sepedi code-switched tweets



Class	Number	%
POS	1,498	52.3
NEG	852	29.8
NEU	780	27.3
Total	2,862	

Distribution of English-Setswana code-switched tweets

- 28.9% of those tweets contain code-switches of Sepedi and English (11,830 tweets).
- 6.9% of those tweets have code-switches of Setswana and English (2,862 tweets).
- **Linguistic challenges**: spelling errors, local jargon, ambiguities, homographs, and tonal words.
- lack of diacritics.
- The socio-cultural background is necessary to annotate tweets correctly.



Contributions

SAfriSenti — Sentiment corpus for Sepedi, Setswana and English.

Sentiment annotation tool SentiApp.

Sentiment lexicons for Sepedi and Setswana.

Statistical analyses and SAfriSenti's linguistic challenges.

Conclusion & Future Work

- SAfriSenti —a large-scale Twitter-based multilingual sentiment corpus for South African languages in a multilingual setting.

- 36.6% of code-switched tweets demonstrate that SAfriSenti is highly multilingual.

- We described our methods for:
    - tweets annotation which contains tweets collection via Twitter API,
    - text processing and normalisation,
    - removal of short and duplicated tweets,
    - pre-annotation based on keywords and emoticons,
    - and annotation based on strict guidelines.

- In future, we plan to:
    - Optimize our data annotation process with the help of machine learning to reduce the manual annotation effort.
    - 250k tweets per language for collection.

References

- Aguero-Torales, M. M., Abreu Salas, J. I., and Lopez Herrera, A. G. (2021). Deep learning and multilingual sentiment analysis on social media data: An overview. Applied Soft Computing, 107:107373.
- Araujo, M., Reis, J., Pereira, A., and Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In Proceedings of the 31st Annual ACM Symposium on Applied Computing, pages 1140–1145.
- Statista. (2022). African countries with the largest population as of 2020.
- Balahur, A. and Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Computer Speech & Language, 28(1): 56–75.
- Becker, W., Wehrmann, J., Cagnini, H. E., and Barros, R. C. (2017). An efficient deep neural architecture for multilingual sentiment analysis in Twitter. In The Thirtieth International Flairs Conference, pages 246–251.
- Ohman, E. (2020). Challenges in annotation: Annotator experiences from a crowdsourced emotion annotation task. In Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, number 2612 in CEUR workshop proceedings, pages 293–301

Thank You

Questions?