Proceedings of SIGUL2022 @LREC2022, pages 70–77
Marseille, 24-25 June 2022

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

Koena Ronny Mabokela, Tim Schlippe
University of Johannesburg, South Africa
IU International University of Applied Sciences, Germany
krmabokela@gmail.com, tim.schlippe@iu.org

Abstract

Multilingual sentiment analysis is a process of detecting and classifying sentiment based on textual information written in multiple languages. There has been tremendous research advancement on high-resourced languages such as English. However, progress on under-resourced languages remains underrepresented with limited opportunities for further development of natural language processing (NLP) technologies. Sentiment analysis (SA) for under-resourced language still is a skewed research area. Although, there are some considerable efforts in emerging African countries to develop such resources for under-resourced languages, languages such as indigenous South African languages still suffer from a lack of datasets. To the best of our knowledge, there is currently no dataset dedicated to SA research for South African languages in a multilingual context, i.e. comments are in different languages and may contain code-switching. In this paper, we present the first subset of the multilingual sentiment corpus *SAfriSenti* for the three most widely spoken languages in South Africa—*English*, *Sepedi (i.e. Northern Sotho)*, and *Setswana*. This subset consists of over 40,000 annotated tweets in all the three languages including even 36.6% of code-switched texts. We present data collection, cleaning and annotation strategies that were followed to curate the dataset for these languages. Furthermore, we describe how we developed language-specific sentiment lexicons, morpheme-based sentiment taggers, conduct linguistic analyses and present possible solutions for the challenges of this sentiment dataset. We will release the dataset and sentiment lexicons to the research communities to advance the NLP research of under-resourced languages.

Keywords: Multilingual, Sentiment analysis, Under-resourced languages, Code-switching, Sepedi, Setswana, South African languages

1. Introduction

Detecting sentiments or emotions from language has been a significant area of research in natural language processing (NLP) for the past decades (Medhat et al., 2014; Wankhade et al., 2022). Sentiment analysis (SA) is concerned with detecting and categorising emotions from textual information (Pang et al., 2002). SA has garnered a lot of research attention which may be attributed to its numerous essential NLP applications. Recently, SA has given birth to multilingual SA due to the rapid use of mixture of languages on various social media platforms (Balahur and Turchi, 2014). Multilingual SA aims to detect and recognise the sentiment of textual information written in more than one language. It is an emerging NLP research area with promising progress on high-resourced languages, i.e., English and Chinese (Ruder, 2020). However, the same cannot be said for languages with limited resource data which continue to remain highly underrepresented. In addition, the lack of resources poses a significant challenge for language-specific services in developing countries (Dashtipour et al., 2016; Lo et al., 2017).

In context, under-resourced languages are in desperate need of data, digital tools, and resources to overcome the resource barrier and enable NLP to deliver more widespread benefits (Ruder, 2020). Developing such language technologies and curated datasets for these under-resourced languages opens a considerable amount of economic perspectives and it is crucial for data availability and training of NLP applications (Marivate et al., 2020). Past research has yielded relatively limited insights into the relationship between socio-cultural factors, multicultural factors and NLP for under-resourced languages (Lo et al., 2017). However, recent research suggests that socio-cultural factors and multicultural diversity impede NLP for under-resourced languages, possibly leading to economic disparities in many multilingual communities (Weidinger et al., 2021)

With at least 7,000 spoken languages world-wide (Ruder, 2020), not many are represented on the internet, including over 2,000 native languages in Africa[1]. South Africa is with over 60 million people, 11 official spoken languages and over 40 dialects not only the sixth African country with the largest population (Statista, 2022). It is also the most multilingual and multicultural society where most native speakers are fluent in at least two languages. 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 (Lama, 2020). However, there has been no SA research at all for the indigenous South African languages, especially

[1] https://www.ethnologue.com

not for Twitter. Therefore, a tremendous effort to create digital resources for such under-resourced languages is necessary for future digital language technologies.

In this paper, we present a subset of *SAfriSenti*—our large-scale multilingual Twitter sentiment corpus for the South African languages English, Sepedi, and Setswana. It is to date the largest annotated sentiment dataset combining Sepedi and Setswana as under-resourced languages and English. We further present strategies to perform data collection using a multi-distant supervision approach, data preprocessing and data annotation which can be extended to other languages with limited data. Particularly, we describe our solutions for the missing support of Sepedi and Setswana in the Twitter API. In more detail in this paper, we offer the following contributions:

- We present a subset of our large-scale multilingual sentiment dataset for South African languages *SAfriSenti*. This subset contains Sepedi, Setswana, and English in a multilingual setting.
- We present our sentiment annotation tool *Senti-App* which allows the combination of automatic sentiment labelling and human annotation.
- We leverage the commonly used English sentiment lexicons AFFIN, NRC and VADER (Hutto and Gilbert, 2015; Nielsen, 2011) to built dedicated sentiment lexicons for Sepedi and Setswana.
- We present statistical analyses of *SAfriSenti*'s subset as well as linguistic challenges. Additionally, we describe how we plan to resolve the discovered challenges.

This paper is organised as follows: Section 2 will describe related work. In section 3, we will discuss our data collection, quality assurance and data annotation methods. In Section 4, we will present statistics of the final high-quality subset. Finally, we provide a conclusion of our research work in section 5 and offer suggestion for future work.

2. Related Studies

The research interest to solve the challenges of under-resourced languages has increased (Aguero-Torales et al., 2021; Wankhade et al., 2022). SA for monolingual, code-switched and multilingual comments on under-resourced languages has been studied only for a few African languages, e.g. several Nigerian languages (Hassan Muhammad et al., 2022), Swahili (Martin et al., 2021) and Bambara (Konate and Du, 2018). SA studies on under-resourced languages used datasets which consist of movie reviews, Amazon reviews, YouTube comments, tweets and Facebook comments (Balahur and Turchi, 2012b; Pan et al., 2011; Pak and Paroubek, 2010). As these datasets contain comments in multiple languages, they are interesting for the multilingual SA research (Vilares et al., 2016; Araujo et al., 2016; Can et al., 2018).

Several researchers investigated cross-lingual methods to solve the challenges of under-resourced languages by utilising language knowledge from high-resourced languages like English (Araújo et al., 2020; Balahur and Turchi, 2014; Can et al., 2018; Vilares et al., 2017). Notably, they frequently translate the comments from the original under-resourced language to English. This enables SA to conduct its classification task with high-performing models that have been trained with a large number of English resources. However, even this approach was successful for the high-resourced languages Russian, German and Spanish (Shalunts et al., 2016), (Ghafoor et al., 2021) report that translation from English to German, Urdu, and Hindi had a bad impact on SA performance. Additionally, (Becker et al., 2017) state that SA is dependent on MT quality in cross-lingual SA. According to (Ghafoor et al., 2021), there was a 2-3% SA performance decrease from English to under-resourced languages with help of MT compared to human translation.

Due to the mentioned MT performance issues, there are monolingual SA approaches for under-resourced languages. For these approaches, data in the target language such as sentiment lexica or labelled comments are required. Usually those data are created in a semi-automatic way, i.e. first machine-translated then manually corrected. (Mihalcea et al., 2012) constructed a Romanian sentiment lexicon (also denoted as subjectivity lexicon) with the help of an English sentiment lexicon and an English-Romanian dictionary. Additionally, (Balahur and Turchi, 2012a) generate SA datasets in the target language with the help of MT. They even claim that SA may be done on these translated data without any significant loss of accuracy. However, (Deriu et al., 2017) demonstrated that for German, English, and Italian there is an SA performance degradation after translating the resources into the target language. Previous studies investigated data collection strategies for under-resourced languages on Twitter (Pak and Paroubek, 2010; Vosoughi et al., 2016). The methods focus on labelling only two sentiment classes —positive and negative. Meanwhile other research work has explored strategies to label three sentiment classes in Twitter—positive, neutral, and negative —using human annotators (Vilares et al., 2016; Pak and Paroubek, 2010; Pang et al., 2002; Nakov et al., 2019). Despite the attempt to automate the data labelling process (Kranjc et al., 2015), the hand-crafted annotation is to date the most preferred method of data labelling in many NLP tasks (Muhammad et al., 2022). However, manual annotation presents challenges and it is deemed an expensive process. Notably, the work presented in (Jamatia et al., 2020; Gupta et al., 2021) employed manually annotated tweets, while other studies focus on automated data labelling solutions (Kranjc et al., 2015). (Vosoughi et al., 2016) investigated var-

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

ious pipelines to collect data on Twitter using distant supervised learning. In this approach, they use positive and negative emoticons as indicators to annotate tweets. (Go et al., 2009) explore distant supervision methods to label millions of tweets using positive and negative search terms (i.e. term queries) in the Twitter API and emoticons to pre-classify the tweets. (Vilares et al., 2016) also investigate *SentiStrength* scores to label an English-Spanish code-switching Twitter corpus. *SentiStrength* is an online SA system available for a few languages (Thelwall et al., 2011).

Compared to (Cliche, 2017; Jamatia et al., 2020; Vilares et al., 2016), we also investigate distant supervised annotation methods with the help of emoticons, search terms and sentiment lexicons. In addition to these methods, to make sure that we only collect tweets in our target languages, we leverage from Twitter's geolocation functionality and language identification based on word frequencies. Finally, the dataset is double-checked by human annotators. Despite Afrikaans (Kotzé and Senekal, 2018) and English, no other South African language has been investigated for SA to the best of our knowledge. We are the first to develop SA resources and systems for Sepedi and Setswana in a multilingual environment. In the next section, we will discuss our data collection strategies for Sepedi, Setswana and English.

3. Data Collection and Preprocessing

In this section, we will first describe the data collection strategies with the help of the Twitter API. Then we will present our methods for text preprocessing and normalization. Table 1 shows an extract from the dataset with examples of tweets in Sepedi, Setswana and English. It further contains English (marked in blue) and Sepedi (marked in blue) code-switched tweets.

3.1. Twitter Data Collection

Twitter provides easy access to a large amount of public user-generated text. It is used by different people to express their opinion about different topics (Pak and Paroubek, 2010). The Twitter API has evolved over time by introducing a new degree of access to enable developers and academic researchers to investigate the public comments for various NLP tasks[2]. We requested the permission to access the Twitter API by explaining our use-cases, agreeing on theirs terms of usage and policies. Our goal was to collect:

1. tweets only from the target languages.
2. trending tweets.
3. tweets with emotions.

We collected the tweets using the Twitter API for Academic Research[3]. For some languages, this API provides a functionality to only collect tweets in one specific language. However, this is not supported for Sepedi and Setswana. Consequently, we implemented word frequency based language identifications to only collect tweets in our target languages. To collect the trending tweets, we requested several native speakers to provide the trending search keywords and hashtags on a website. With the help of the Twitter API we only collected tweets which contain emoticons to ensure that those tweets contain emotions.

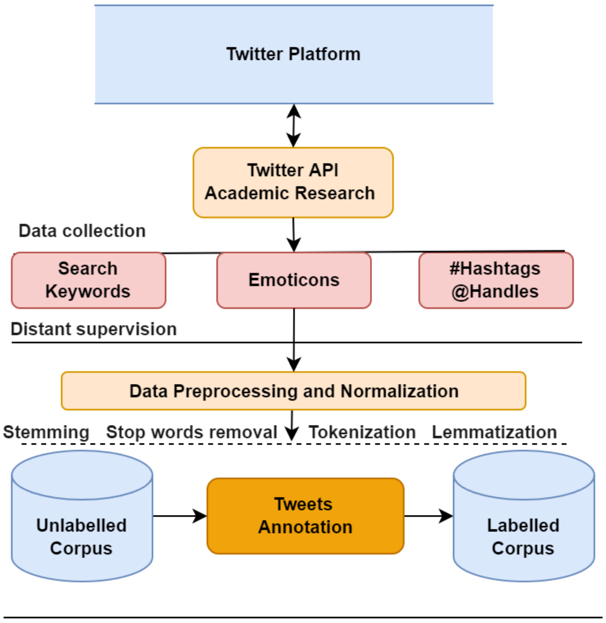

Figure 1: Data collection, cleaning and annotation

Figure 1 summarises these methods for data collection, plus our methods for cleaning and annotation, which we will describe in the upcoming subsection.

[2] https://developer.twitter.com/en/docs/twitter-api

[3] https://developer.twitter.com/en/products/twitter-api/academic-research

3.2. Text Preprocessing and Normalisation

We performed preprocessing, normalisation, lammentazation and tokenization on each tweet as used in (Pang et al., 2002; Pak and Paroubek, 2010). Each tweet was preprocessed in the following steps:

1. We remove very short tweets and duplicated tweets.
2. With the help of the @ symbol, we substitute people's and company's names for the purposes of data protection.
3. We remove punctuations, URLs and the # symbol.
4. We remove characters that appear more than twice (e.g., **`Loooool or Whaaaaaat`** and **`ngwanaaaaaka`** is replaced with **`Lol`** or **`What`** and **`ngwanaka`**).
5. We substitute abbreviations by their long form.
6. We set all words to lowercase, remove unnecessary white spaces and tokenize the tweets using the NLTK tokenizer (Bird and Loper, 2004).

In the next subsections, we will describe the preprocessing steps 1 and 3 in more detail.

3.3. Removal of short and duplicated tweets

We remove duplicated tweets in step 1 because they do not contain any additional information. We handle retweets and quote tweets with an **`@RT`** tag in the following way: We remove tweets which only contain a retweet. In those tweets which contain a quote of another tweet, we only keep the text which is new since we think it contains valuable information. To make sure that we get useful information in the tweets, we remove tweets with less than 5 word tokens.

4. The SAfriSenti Corpus

After we have described the text preprocessing steps, we will depict how we labelled the remaining English, Sepedi and Setswana tweets.

4.1. Pre-annotations

As mentioned in section 2 and recommended by (Go et al., 2009; Vosoughi et al., 2016), we used emoticons as distantly supervised method to pre-classify tweets as *positive*, *neutral* or *negative*. For this, we derived our initial sentiment classes from emoticons representing happy, smile, love, angry and sad as in (Pak and Paroubek, 2010; Nakov et al., 2019). In some tweets, users express their opinions using multiple unrelated emoticons which makes the pre-classification difficult. Consequently, we additionally checked the tweets for words in a sentiment lexicon which will be described in section 5.1. Then human experts verified the pre-classified tweets.

4.2. Annotation Guidelines

We defined strict annotation guidelines which all annotators have to follow in the decision to classify the tweets into *positive* (POS), *neutral* (NEU) and *negative* (NEG) as in (Turney, 2002; Öhman, 2020). We adopted our guidelines from (Mohammad, 2016) and consulted 3 language experts for each language to double-check our guidelines. The annotation guidelines for labelling our sentiment classes are summarised as follows:

- **Positive Sentiment (POS)** - This happens when a tweet expresses a favorable viewpoint, expression of support, appreciation, positive attitude, forgiveness, encouragement, success, cherish or pleasant emotional state.
- **Negative Sentiment (NEG)** - This happens when a tweet contains negative words, such as criticism, judgment, negative attitude, doubting validity/competence, failure or negative emotion.
- **Neutral Sentiment (NEU)** - This happens when a tweet does not directly or indirectly imply any positive or negative words. Typically, these are factual tweets such as reports or general statements.
- **Positive and Negative Sentiment** - This happens when a tweet expresses a positive language in part and negative language in part. For all three languages, these tweets are classified as positive or negative based on a score computed from individual scores in the corresponding sentiment lexicon. For Sepedi and Setswana we additionally apply language-specific morphological rules which will be explained in section 5.2.

Our annotation guidelines further contain the labelling of tweets with code-switches as well as tweets with no sentiments as the following classes:

- **Mixed Language (MIX)** - This happens when a tweet contains text from several languages (i.e. code-switched text).
- **None Sentiment (NOS)** - This happens when a tweet has no indication of the sentiment due to lack of context, e.g. in proverbs, idioms, or sarcasm.

4.3. Annotator's background and training

We recruited 3 native speakers with technical and linguistic background for each language as annotators. To facilitate the labelling process, we developed *SentiApp*, an online platform for organising and annotating the tweets. In a training session, we informed our annotators about our annotation guidelines and demonstrated the use of *SentiApp*. Then they were first asked to annotate 150 tweets. After their annotation process, they received our feedback to improve quality for upcoming tweets as recommended by (Öhman, 2020).

4.4. Annotation Process

After our training session, for organisational reasons, every time our annotators labelled batches with 1,000 tweets in our *SentiApp*. All three representatives of a language always worked on the same batch to be able to compare the resulting labels. In case of disagreement, the final label is determined by a majority voting. We also validated instances where annotators provided the labels NOS and MIX. As in (Muhammad et al., 2022), tweets with NOS are excluded from the dataset since they do not contain any sentiment. In total, the annotators labelled 25,947 monolingual tweets and 14,692 tweets which contain code-switches.

To determine the final label, we used a majority voting approach together with the proposed strategies of (Davani et al., 2021) which deals with the following 4 cases:

- **Three-way disagreement**—This happens when all 3 annotators disagree on a label. For example, if a tweet is labelled as NEG, NEU and POS. In this case, the annotators double-check these tweets or in case of remaining disagreement we discard this tweet.

- **Three-way agreement**—This happens when all 3 annotators agree on a label. For example, if a tweet is labelled as NEG by all 3 annotators, then it is NEG.

- **Two-way partial disagreement**—This happens when 2 annotators agree on a label but the third annotator chooses the label NEU. For example, if a tweet is labelled by 2 annotators as POS and by the other annotator as NEU, the final label is POS.

- **Two-way disagreement**—This happens when 2 annotators agree on a label but the third annotator chooses another label which is not NEU. For example, if a tweet is labelled by 2 annotators as POS and by the other annotator as NEG, the final label is POS.

4.5. Data Statistics

In total, we collected over 250,000 tweets for our 3 languages. However, in this paper, we report only the annotated subset of over 40,000 tweets. Tables 2 to 6 show an overview of the monolingual and code-switched tweets in this annotated subset. The monolingual tweets cover 63.4% (25,947 tweets). As demonstrated in Tables 5 and 6, our subset consists of a large number of code-switched tweets (14,692 tweets). 28.9% of those tweets contain code-switches of Sepedi and English (11,830 tweets). 6.9% of those tweets contain code-switches of Setswana and English (2,862 tweets). Sepedi and Setswana share some common words since the languages are closely-related.

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

Table 2: Distribution of Sepedi tweets

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

Table 3: Distribution of Setswana tweets

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

Table 4: Distribution of English tweets

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

Table 5: 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	

Table 6: Distribution of English-Setswana code-switched tweets

4.6. Linguistic Challenges

Sepedi has diacritics, while Setswana does not have any diacritics. In some Sepedi tweets the diacritics are expressed with Roman characters, e.g. š is replaced by sh, ch or x due to the use of English keyboards. These replacements of Sepedi diacritics sometimes leads to character strings which are very similar to Setswana words. Linguistic challenges, particularly evident in Sepedi, are that some tweets contain spelling errors, local jargon, ambiguities, homographs, and tonal words. Tones in Sepedi give meaning to words, particularly those words which have the same orthographic representation. For example, the word *noka*—depending on the context and tone, means "river" or "waist": The sentence "*ke tlo boya gae ge noka ke sena meetse*" meaning "*I will come back when the river has no water*" has a positive sentiment but the sentence "*o dula o* *bolaya ke noka buti wa tšwafa*" meaning "*My brother, you are always complaining about your waist, you are lazy*" has a negative meaning. In addition to the linguistic challenges, we encountered that knowledge of socio-cultural background is necessary to correctly label some tweets. We assume that this required additional socio-cultural knowledge could be a challenge for automatic SA systems. Furthermore, some tweets in *SAfriSenti* contain emoticons. On the one hand, these emoticons can be an indicator for the correct sentiment class. On the other hand, for tweets which contain multiple emoticons that expresses contradictory emotions, finding the correct sentiment class can be a challenging task.

5. Additional Resources

In addition to the *SAfriSenti* Corpus, we created sentiment lexicons and morpheme-based sentiment taggers.

5.1. Sentiment Lexicons

Figure 2 depicts the framework for building our sentiment lexicons:

1. Our annotators mark the sentiment-bearing words in each tweet which are only contained in positive and negative tweets.

2. We built a wordlist with the sentiment-bearing words and delete duplicates.

3. To each word we add the sentiment labels of the corresponding tweet and receive a sentiment lexicon.

4. As in (Nielsen, 2011), we let our annotators score the sentiment strength of positive words in a range between +1 (very weak) and +5 (very strong) and the strength of negative words in a range between −1 (very weak) and −5 (very strong).

5. We translate the words in the sentiment lexicon.

6. We merge our sentiment lexicon with translated versions of the well-known English sentiment lexicons AFFIN and VADER (Hutto and Gilbert, 2015; Nielsen, 2011).

To translate between Sepedi and English, we first used the Google Translate API and between Setswana and English, we used the Autshumato MT Web Service[4]. Autshumato is an open-source translation system, which was developed by *Centre for Text Technology* (CTexT) at the North-West University. Then our annotaters double-checked and corrected the translations.

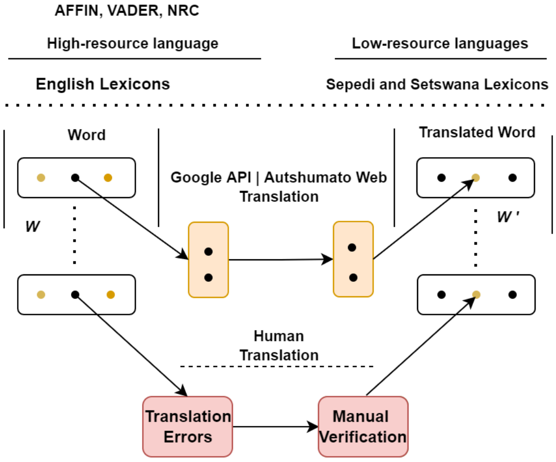

Figure 2: Developing sentiment lexicons.

5.2. Sentiment Taggers

As an additional information source, we looked into the morpheme level of Sepedi and Setwana since those language often contain morphemes that indicate the mood. Consequently, we also developed sentiment taggers for Sepedi and Setswana that first split individual words into morphemes and then label those words based on specific morphemes that indicate positive or negative moods. Examples for Sepedi morphemes which indicate a negative mood are: `/ke be ke sa/ and /ba be ba sa/`. Examples for Sepedi morphemes which indicate a positive mood are: `/ke be ke/ or /ba be ba/`. In the future we plan to investigate the application of our sentiment taggers as additional source for the sentiment classification.

6. Conclusion and Future Work

This paper presented a subset of *SAfriSenti*—a large-scale Twitter-based multilingual sentiment corpus for South African languages in a multilingual setting. We are the first who collected the under-resourced languages Sepedi and Setswana in this corpus. 36.6% of code-switched tweets demonstrate that *SAfriSenti* is highly multilingual. We described our methods for tweets annotation which contain tweets collection via the Twitter API, text processing and normalisation, removal of short and duplicated tweets, pre-annotation based on emoticons, and annotation based on strict guidelines. In addition, we discussed the challenges and mitigation of our data collection process. Additionally, we created sentiment lexicons for Sepedi and Setswana as well as implemented sentiment taggers which use morphemes to indicate the sentiment class.

In the future, we plan to optimize our data annotation process with the help of machine learning to reduce the manual annotation effort iteratively, similar to (Schlippe et al., 2012). Further our goal is to expand *SAfriSenti* with more African under-resourced languages to release a large-scale Twitter based mul-

[4] https://mt.nwu.ac.za

tilingual sentiment corpus for SA to the NLP research community. Moreover, we will use *SAfriSenti* to investigate and compare different approaches for SA. Since we encountered that knowledge of cultural background is necessary to correctly label some tweets, we will analyze methods to leverage socio-cultural information into SA systems.

7. Acknowledgments

We acknowledge the National Research Foundation (NRF) for sponsorship. We also want to express our gratitude to our annotators and facilitators. Finally, we thank the reviewers for their valuable feedback.

8. 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.

Araújo, M., Pereira, A., and Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, 512:1078–1102.

Balahur, A. and Turchi, M. (2012a). Comparative experiments for multilingual sentiment analysis using machine translation. In *SDAD@ ECML/PKDD*, pages 75–86.

Balahur, A. and Turchi, M. (2012b). Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 52–60. Association for Computational Linguistics.

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.

Bird, S. and Loper, E. (2004). NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July. Association for Computational Linguistics.

Can, E. F., Ezen-Can, A., and Can, F. (2018). Multilingual sentiment analysis: An RNN-based framework for limited data. *arXiv preprint arXiv:1806.04511*.

Cliche, M. (2017). BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs. *arXiv preprint arXiv:1704.06125*.

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., and Zhou, Q. (2016). Multilingual sentiment analysis: State of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771.

Davani, A. M., Díaz, M., and Prabhakaran, V. (2021). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *CoRR*, abs/2110.05719.

Deriu, J., Lucchi, A., Luca, V. D., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., and Jaggi, M. (2017). Leveraging large amounts of weakly supervised data for multi-language sentiment classification. *Proceedings of the 26th International Conference on World Wide Web*, pages 1045–1052.

Ghafoor, A., Imran, A., Daudpota, S., Kastrati, Z., Abdullah, Batra, R., and Wani, M. (2021). The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9:124478 – 124490. Cited by: 0; All Open Access, Gold Open Access, Green Open Access.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*, 150, 01.

Gupta, A., Menghani, S., Rallabandi, S. K., and Black, A. W. (2021). Unconscious self-training for sentiment analysis of code-linked data. *ArXiv*, abs / 2103.14797.

Hassan Muhammad, S., Ifeoluwa Adelani, D., Ruder, S., Said Ahmad, I., Abdulmumin, I., Shehu Bello, B., Choudhury, M., Chinenye Emezue, C., Salahudeen Abdullahi, S., Aremu, A., Jeorge, A., and Brazdil, P. (2022). NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis. *arXiv e-prints*, page arXiv:2201.08277, January.

Hutto, C. and Gilbert, E. (2015). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, pages 218–225, 01.

Jamatia, A., Swamy, S. D., Gamback, B., Das, A., and Debbarma, S. (2020). Deep Learning Based Sentiment Analysis in a Code-Mixed English-Hindi and English-Bengali Social Media Corpus. *International Journal on Artificial Intelligence Tools*, 29.

Konate, A. and Du, R. (2018). Sentiment Analysis of Code-Mixed Bambara-French Social Media Text Using Deep Learning Techniques. *Wuhan University Journal of Natural Sciences*, 23:237–243, 06.

Kotzé, E. and Senekal, B. (2018). Employing sentiment analysis for gauging perceptions of minorities in multicultural societies: An analysis of Twitter feeds on the Afrikaner community of Orania in South Africa. *TD: The Journal for Transdisciplinary Research in Southern Africa*, 14(1):1–11.

Kranjc, J., Smailovic, J., Podpecan, V., Grcar, M., Znidari, M., and Lavrac, N. (2015). Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the clowdflows platform. *Inf. Process. Manag.*, 51:187–203.

Lama. (2020). Talkwalker: Social Media Statistics and Usage in South Africa.

Lo, S. L., Cambria, E., Chiong, R., and Cornforth, D. (2017). Multilingual sentiment analysis: from formal to informal and scarce resource languages. *Artificial Intelligence Review*, 48(4):499–527.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T. B., Mokoena, R., and Modupe, A. (2020). Low resource language dataset creation, curation and classification: Setswana and Sepedi - Extended Abstract. *CoRR*, abs/2004.13842.

Martin, G. L., Mswahili, M. E., and Jeong, Y.-S. (2021). Sentiment Classification in Swahili Language Using Multilingual BERT. *African NLP Workshop, EACL 2021*, abs/2104.09006.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Mihalcea, R., Banea, C., and Wiebe, J. (2012). Multilingual subjectivity and sentiment analysis. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, page 4, Jeju Island, Korea, July. Association for Computational Linguistics.

Mohammad, S. (2016). A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 174–179, San Diego, California, June. Association for Computational Linguistics.

Muhammad, S. H., Adelani, D. I., Ruder, S., Ahmad, I. S., Abdulmumin, I., Bello, B. S., Choudhury, M., Emezue, C. C., Abdullahi, S. S., Aremu, A., Jeorge, A., and Brazdil, P. (2022). NaijaSenti: A Nigerian Sentiment Corpus for Multilingual Sentiment Analysis.

Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., and Stoyanov, V. (2019). SemEval-2016 task 4: Sentiment analysis in Twitter. *arXiv preprint arXiv:1912.01973*.

Nielsen, F. Å. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903.

Öhman, 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, International. CEUR Workshop Proceedings.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, pages 1320–1326.

Pan, J., Xue, G.-R., Yu, Y., and Wang, Y. (2011). Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 289–300. Springer.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Ruder, S. (2020). Why you should do NLP beyond English. `http://ruder.io/nlp-beyond-english`.

Schlippe, T., Ochs, S., and Schultz, T. (2012). Grapheme-to-phoneme model generation for Indo-European languages. In *The 37th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, 25-30 March.

Shalunts, G., Backfried, G., and Commeignes, N. (2016). The impact of machine translation on sentiment analysis. In *The Fifth International Conference on Data Analytics*.

Statista. (2022). African countries with the largest population as of 2020.

Thelwall, M. A., Buckley, K., and Paltoglou, G. (2011). Sentiment in Twitter events. *J. Assoc. Inf. Sci. Technol.*, 62:406–418.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. (2016). EN-ES-CS: An English-Spanish code-switching Twitter corpus for multilingual sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4149–4153.

Vilares, D., Alonso, M. A., and Gómez-Rodríguez, C. (2017). Supervised sentiment analysis in multilingual environments. *Information Processing & Management*, 53(3):595–607.

Vosoughi, S., Zhou, H., and Roy, D. (2016). Enhanced Twitter sentiment classification using contextual information. *CoRR*, abs/1605.05195.

Wankhade, M., Rao, A., and Kulkarni, C. (2022). A Survey on Sentiment Analysis Methods, Applications, and Challenges. *Artificial Intelligence Review*, pages 1–50, 02.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W. S., Legassick, S., Irving, G., and Gabriel, I. (2021). Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.