2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)

Sentiment Analysis for Shona

Barlette Makuwe
IU International University of Applied Sciences
Germany
barmakuwe@gmail.com

Koena Ronny Mabokela
University of Johannesburg
South Africa
krmabokela@gmail.com

Tim Schlippe
IU International University of Applied Sciences
Germany
tim.schlippe@iu.org

***Abstract*—No sentiment analysis system existed for Shona yet—even though it is a Bantu language spoken by nearly 17 million people. Consequently, we collected *ShonaSenti*—a new corpus of 16,000 tweets in Shona covering 8 different topics. In this paper, we describe our distant supervised labelling strategies to support the annotators to categorize the collected tweets into the 5 sentiment classes of *very negative*, *negative*, *neutral*, *positive* and *very positive*. Moreover, we leveraged the Shona sentiment analysis corpus to develop first mono-lingual and cross-lingual sentiment analysis systems for Shona. Our best sentiment analysis systems are cross-lingual and mono-lingual Transformer-based systems which achieve accuracies of 84% on 3 sentiment classes and 70% on 5 sentiment classes.**

***Index Terms*—sentiment analysis, Shona, corpus, distant supervision, low-resource language, African language**

I. Introduction

Sentiment analysis is the process of automatically detecting a sentiment from textual information and then classifying the information into classes such as *negative*, *neutral* or *positive* [1]. Its application draws attention not only in business environments [2] but also in other areas, like medicine [3], education [4] [5] and AI for Social Good [6]. In natural language processing (NLP) research, Twitter has proven to be a good source of sentiment-related textual data [7]–[10].

Many of the sentiment analysis applications are developed for English. But English is only spoken by 19% of the world population [11]. Consequently, more and more NLP corpora and systems are developed for the languages spoken in the economically stronger countries. However, there are still many thousands of so-called low-resource languages in the economically weaker countries. Africa and India alone have approximately 2,000 low-resource languages with more than 2.5 billion people [12] [13]. In order to give them the chance of benefiting from the economic, medical and social advantages of sentiment analysis, it makes sense to develop data and systems for these languages as well.

Shona is such a low-resource language. The language is the mother tongue of 75% of Zimbabwe's population [14]. Further Shona-speaking populations are located in the nearby nations like southern Zambia, Botswana, Mozambique and South Africa, but also in United Kingdom [15]. No sentiment analysis system exists for Shona yet, even though it is a Bantu language which is spoken by 16.6 million people [16].

Consequently, our contributions are:

- With the help of Twitter's Academic API, we collected a new corpus of nearly 16,000 tweets in Shona covering 8 topics—the *ShonaSenti* corpus. To contribute to the improvement of low-resource languages, we share the corpus with the research community[1].
- Using our labelling strategy, we had native speakers manually categorize the collected tweets into the 3 classes of *negative*, *neutral*, and *positive*.
- To allow a more detailed sentiment analysis, we also had the native speakers manually categorize the collected tweets into the 5 classes of *very negative*, *negative*, *neutral*, *positive*, and *very positive*.
- We leveraged the Shona sentiment analysis corpus to develop the first mono-lingual and cross-lingual sentiment analysis systems for Shona.

In the next section we will look at related work. In Section 3 we will present our data collection and labeling strategies together with *ShonaSenti* corpus' statistics. Section 4 will demonstrate our experiments with the mono-lingual and cross-lingual sentiment analysis systems. In Section 5, we will summarize our work and propose possible future work.

II. Related Work

While sentiment analysis resulted in many successful applications in different fields like business [2] [17] and medicine [3], it has also been the subject of research in education [4] [5]. Furthermore, it has proven helpful in the field of AI in Social Good. For example, [6] applied cross-lingual sentiment analysis on South African tweets to detect social challenges described in English, Sepedi (i.e. Northern Sotho) and Setswana tweets. The classification of sentiments is a discrete approximation of a continuous scale on which sentiments are located. Therefore, besides the traditional classification into the 3 classes *negative*, *neutral* and *positive*, there are first approaches of sentiment analysis systems that use more than 3 classes for a more detailed mapping of sentiments, such as [18].

Sentiment analysis corpora are usually retrieved from microblogging services like Twitter since these sources share situational information, cover a lot of topics and contain negative, neutral and positive tweets [19] [20]. Several studies investigated different data collection methods for tweets [19]

[1] https://github.com/barletteM/Shona-Sentiment-Analysis

979-8-3503-2743-4/ 23/$31.00

[7] [21] [9] [20]. [19] explored methods to collect millions of annotated tweets from various places, hours, and writers. Other studies used emoticons and keywords [7] [9] to extract and build Twitter-based corpora via distant supervision. Furthermore, to ensure correct labelling, [8] let the tweets be labelled by three annotators following the *SentiStrength* strategy [22]. [23] used distant supervised methods with emoticons and keywords together with a word frequency-based language identification to collect tweets in Sepedi, Setswana and English.

For automatic sentiment analysis, different machine learning algorithms like support vector machines, decision trees, random forests, multilayer perceptrons and long short-term memories were analysed [24]–[27]. [5] demonstrated that the Transformer models BERT (Bidirectional Encoder Representations from Transformers) [28] and RoBERTa (Robustly Optimized BERT Pretraining Approach) [29] usually outperform the other machine learning algorithms. Lexicon-based approaches were also investigated, e.g. in [30] [31], but machine learning algorithms usually perform better than the lexicon-based approaches. AfroLM is a Transformer model pre-trained on 23 African languages [32]. It is the only NLP model which is able to deal with Shona. Consequently, we used AfroLM for our mono-lingual sentiment analysis systems, even though it was pre-trained on only 32.8 MB of Shona text [32].

To solve the problems of low-resource languages, some researchers propose cross-lingual NLP approaches. Thus is it possible to benefit from rich-resource languages like English [24] [33] [22] [34]. For sentiment analysis, they usually translate the comments from the original low-resource language to English. This allows to do the classification task of sentiment analysis with well-performing models trained with a lot of English resources. Therefore, we used English BERT which was pre-trained with the BooksCorpus (800M words) [35] and English Wikipedia (2,500M words) [28] for the cross-lingual experiments.

To collect our Shona sentiment analysis corpus, we applied a labelling strategy which is similar to the *SentiStrength* strategy [22]. But to accelerate the labelling process, we additionally leveraged distant supervised methods with emoticons and keywords as described in [23]. Then we investigated mono-lingual and cross-lingual sentiment analysis systems for Shona.

III. ShonaSenti: The Shona Sentiment Analysis Corpus

In this section we will describe our collection of Shona tweets and our labelling strategies which include sentiment-lexicon and emoticon-based distant supervision approaches. Then we will present statistics of our collected corpus *ShonaSenti*.

A. Collection of the Tweets

Our goal was to collect text data for sentiment analysis which covers a variety of topics. Since tweets provide a reflection of society's topics [36], we used the Essential Twitter API package for our text data collection. Additionally, to cover our 8 topics, we searched for tweets with topic-specific search terms. To increase the chance to obtain tweets in Shona, we limited the search with a geocode representing the location of Zimbabwe and a radius. Specifying the geolocation of tweets was needful as some of the search terms we used are part of other Bantu languages across the African continent as well as other languages are spoken across the globe. Zimbabwe spans an area of 390,757 square kilometres [37]. Therefore, a search radius of 625 kilometres around the centre of Zimbabwe was sufficient to exclude neighbouring countries such as South Africa, Botswana, Zambia, etc., where other languages are spoken that may contain the same words as our search terms.

With the geolocation-based search, we collected about 21,000 tweets from August to November 2022. To make sure that we kept only Shona tweets for our corpus, we compared the words in each tweet with the most frequent words in Shona and the most frequent words in English. As demonstrated in Table I, our word frequency-based language identification revealed that 75.9% of the tweets contain pure Shona, 10.4% Shona-English code-switching, 13.2% pure English, and 0.5% other languages such as Ndebele, Sotho, or Tswana. Since our goal was to collect pure Shona tweets and not to tackle the challenges of code-switching and other languages in this study, we discarded the tweets containing the other languages. 15,960 tweets remained which contained pure Shona.

Language	**Frequency**	**Percentage**
Shona	15,960	75.9%
English	2,776	13.2%
code-switch	2,184	10.4%
others	15,315	0.5%

TABLE I: Languages in the collected tweets.

B. Labelling Strategies

To accelerate the labelling process by reducing the manual effort in a semi-automated manner, we investigated two distant supervision approaches plus their combination.

1) Distant Supervision with Sentiment Lexicon: First, our goal was to automatically pre-label the tweets based on their sentiment-bearing words with the help of a sentiment lexicon (*Sentiment Lexicon*). "A sentiment lexicon is a collection of words associated with their sentiment orientation" [38]. Since no Shona sentiment lexicon exists, we (1) used Google Translate to automatically translate each Shona tweet to English and (2) used [39]'s *Opinion Lexicon*[2] as English sentiment lexicon to compute the coverage of positive and negative words in each tweet. The *Opinion Lexicon* contains two lists of about 6,800 negative and positive words and the lexicon has been successfully used for social media sentiment analysis [39]. Translating the Shona tweets to English instead of the English sentiment lexicon entries to Shona has the advantage that more context is given for the translation which improves translation quality. Then the sentiment label was determined with the following rules:

[2] http://www.cs.uic.edu/ liub/FBS/opinion-lexicon-English.rar

- If more negative words than positive words are found in the tweet, the tweet is labelled as *negative*.
- If more positive words than negative words are found in the tweet, the tweet is labelled as *positive*.
- If the number of positive and negative words are equal in the tweet, the tweet is labelled as *neutral*.
- If no word of the sentiment lexicon is found in the tweet, the tweet is labelled as *unknown (UNK)*.

Sentiment class	**Frequency**	**Percentage**
negative	2,141	13.4%
neutral	4	0.0%
positive	3.569	22.4%
UNK	10,245	64.2%

TABLE II: Sentiment classes with *Sentiment Lexicon*.

Table II shows that with this sentiment lexicon-based distant supervision approach, 35.8% of the tweets could be pre-labelled. The accuracy of this approach is 59% if we take the corresponding labels created and cross-checked by our annotators as reference. This means that with this approach 21% of all 15,960 tweets could be correctly labelled and consequently did not have to be changed in the next step by the annotators. 64.2% of the tweets could not be pre-labelled with this approach since they did not contain sentiment-bearing words from the sentiment lexicon. Furthermore, we see that the sentiment lexicon-based approach has difficulties to find *neutral* tweets, since in only 4 tweets the number of positive and negative words were equal.

2) Distant Supervision with Emoticons: Emoticons express the way people feel and can be useful in determining the sentiment of any subject [40]. Consequently, for our distant supervision approach with emoticons (*Emoticons*), we defined 31 emoticons as indicators for a *negative* classification, 10 emoticons as *neutral* indicators, and 56 emoticons as *positive* indicators. The emoticons were chosen based on personal experience with social media as well as acquaintance with the manner in which the public in Zimbabwe uses emoticons on social media. To be platform-independent, our algorithm finds and compares the emoticons using their ASCII code. The sentiment label was determined as follows:

- If multiple emoticons representing different sentiments are found in the tweet, the tweet is labelled with the sentiment class from which most of the emoticons contained in the tweet are.
- If there is no majority, the tweet is not labelled at this step.

Sentiment class	**Frequency**	**Percentage**
negative	3,300	20.7%
neutral	38	0.2%
positive	345	2.2%
UNK	12,276	76.9%

TABLE III: Sentiment classes with *Emoticons*.

Table III indicates that with this emoticon-based distant supervision approach alone, 23.1% of the tweets could be pre-labelled. The accuracy of this approach is 52.1% if we take the corresponding labels created and cross-checked by our annotators as reference. This means that with this approach only 12.0% of all 15,960 tweets could be correctly labelled and consequently did not have to be changed in the next step by the annotators. 76.9% of the tweets could not be pre-labeld with this approach since they did not contain sentiment-bearing emoticons. As in our sentiment lexicon-based approach, the emoticon-based approach was not successful in finding *neutral* tweets: Only 38 emoticons indicating *neutral* were found in the tweets.

3) Distant Supervision with Sentiment Lexicon and Emoticons: To be able to pre-label more tweets, we combined the sentiment-lexicon-based approach and the emoticon-based approach (*Sentiment Lexicon+Emoticons*) as follows: (1) We applied our sentiment-lexicon-based approach to all tweets. (2) We applied our emoticon-based approach to all tweets. (3 Then we applied the following rules to determine the final sentiment tags:

- **Full-annotation agreement**: If the lexicon- and emoticon-based approaches result in the same sentiment label, the tweet is labelled with this label.
- **Lexicon-override**: If the lexicon- and emoticon-based approaches result in different sentiment labels, the tweet is labelled with the label produced by the lexicon-based approach.
- **Emoticon-control**: If the emoticon-based approach results in a sentiment label but the lexicon-based approach not, the tweet is labelled with the label produced by the emoticon-based approach.

Sentiment class	**Frequency**	**Percentage**
negative	2,931	18.3%
neutral	14	0.1%
positive	3,667	23.0%
UNK	9,347	58.6%

TABLE IV: Sentiment classes with *Sentiment Lexicon+Emoticons*.

Table IV shows that with the combination of the sentiment-lexicon-based approach and the emoticon-based approach, 41.4% of the tweets could be pre-labelled. The accuracy of this approach is 56.9% if we take the corresponding labels created and cross-checked by our annotators as reference. This means that with this approach 23.6% of all 15,960 tweets could be correctly labelled and consequently did not have to be changed in the next step by the annotators. 58.6% of the tweets could not be pre-labelled with this approach since they did not contain sentiment-bearing words from the sentiment lexicon or sentiment-bearing emoticons.

4) Labelling of the Annotators: Each label generated by the distant supervision approach with sentiment lexicon and emoticons was cross-checked by 3 annotators, corrected if necessary, and missing labels were filled in. The annotators are native Shona speakers and have sound educational backgrounds. Strict annotation guidelines for labelling our sentiment classes were drafted to help our annotators understand the importance of the annotation process. Our annotation guidelines were based on the guidelines from [41] [23] for 3 sentiment classes but extended to 5 sentiment classes as follows:

- **Very negative (VNEG)**: A tweet shows an extremely sad, deeply regretful, seriously threatening, swearing, or extremely regretful sentiment—more negative than *NEG*.
- **Negative (NEG)**: A tweet is negatively judgmental, shows criticism, a negative attitude, doubt about validity/competence, failure, dissatisfaction, uses hate speech, or any other negative sentiment, but is less negative than *VNEG*.
- **Neutral (NEU)**: A tweet that does not directly or indirectly imply any positive or negative words, but mainly reflects facts as they appear in reports or general statements, and is neither good nor bad.
- **Positive (POS)**: A tweet shows a favorable viewpoint, an expression of support, appreciation, positive attitude, forgiveness, encouragement, success, pleasant emotional state, or any other positive sentiment, but not as positive as *VPOS*.
- **Very positive (VPOS)**: A tweet shows an overjoyed, excited, thrilled, jubilant, praising, euphoric, elated sentiment or any other expression of being extremely satisfied—more positive than *POS*.

Then we applied the following rules to determine the final sentiment labels:

- **Full agreement**: If all annotators agree on a label, the tweets is labelled with this label.
- **Disagreement**: If not all annotators agree on a label,
(1) map *VPOS* → *POS* and *VNEG* → *NEG*
(2) If 2 labels are the same after the mappings, the tweets is labelled with this label, otherwise the tweets is labelled as *neutral* (*NEU*).

To provide the tweets in chunks and manage our annotators' work, we used the online text annotation tool LightTag[3] [42] which is free for academic research.

The annotator agreement on our 5 classes is listed in Table V. In 65.2% of the tweets, exactly 2 annotators assigned the same label. A complete agreement happened in 34.7% of the tweets. In only 0.1% of the tweets, 3 different labels were assigned. This demonstrates that in 65.3% of the cases, the annotators have different opinions how to classify the tweets.

Agreement	Percentage
all annotators disagree	0.1%
2 annotators agree	65.2%
all 3 annotators agree	34.7%

TABLE V: Annotator agreement for the 5 classes.

C. Statistics

Table VI illustrates the distribution of our collected tweets over topics. The collected 15,960 tweets cover the topics of *sanitation*, *finance*, *agriculture*, *education*, *home affairs*, *communication*, *music*, and *defense*. We see that for some topics such as *sanitation* (3,975 tweets), *finance* (2,244 tweets), *agriculture* (2,190 tweets), and *education* (2,090 tweets), we found more tweets, while for other topics such as *communication* (1,466 tweets), *music* (1,365 tweets), and *defense* (731 tweets) less tweets existed.

Topic	Frequency	Percentage
Sanitation	3,975	24.9%
Finance	2,244	14.1%
Agriculture	2,190	13.7%
Education	2,090	13.1%
Home Affairs	1,899	11.9%
Communication	1,466	9.2%
Music	1,365	8.5%
Defence	731	4.6%

TABLE VI: Topics in our Shona sentiment corpus.

To compare the sentiments of the single topics with each other, we computed the *overall sentiment score* for each topic based on the 3 classes *negative*, *neutral*, and *positive*, as described in [6]:

$$\frac{\#negative * (-1) + 0 * \#neutral * (+1) * \#positive}{\#allsentiments}$$

The *overall sentiment score* lies between -1 and +1, where -1 expresses a completely negative sentiment and +1 a completely positive sentiment. The benefit of this score is that it gives a clear tendency in only one score and makes it easier to compare the topics.

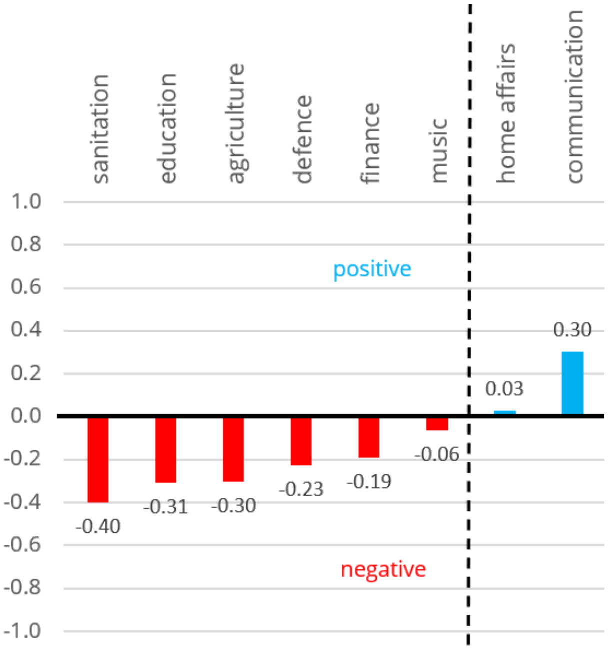

Fig. 1: Overall sentiment scores of the topics.

Figure 1 shows the *overall sentiment score* distribution of our tweets over our 8 topics. We see that the *overall sentiment*

[3] https://www.lighttag.io

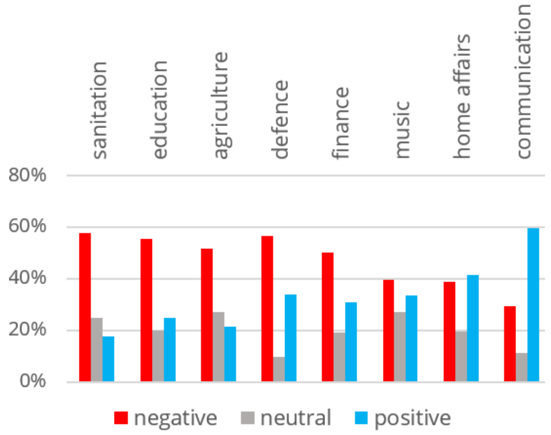

Fig. 2: Overall sentiment scores over the topics.

scores of 8 of the 10 topics are in the negative number range. The topics *sanitation* and *education* are particularly problematic as their scores are less than -0.3, whereas with positive scores the mood regarding *home affairs* and *communication* is rather *positive*. The *overall sentiment score* of all 15,960 tweets in *ShonaSenti* is -0.14, i.e. rather balanced.

Figure 2 shows the sentiment distribution of the 3 classes *negative*, *neutral* and *positive* over our 8 topics. Here, too, we see what the *overall sentiment scores* in Figure 1 have already indicated: The topics *sanitation*, *education*, *agriculture*, *defence*, and *finance* are particularly problematic—more than 50% of the tweets are categorised as *negative*. The mood regarding the topics of *agriculture* and *rural development* is rather *positive*.

To train and evaluate our sentiment analysis systems, we split the corpus into training set and test set as shown in Table VII.

	#sentences	**Percentage**
Train	12,767	80%
Test	3,193	20%

TABLE VII: Distribution of training set and test set.

IV. Experiments and Results

In this section we will present our mono-lingual and cross-lingual systems together with their performances.

A. Overview of the Sentiment Analysis Systems

To analyse sentiment analysis for Shona, we used our training set with 12,767 tweets to train the systems and our test set with 3,193 tweets to evaluate their performances. As visualised in Figure 3, we investigated mono-lingual and cross-lingual sentiment analysis systems to classify the collected tweets into our 3 classes *negative*, *neutral* and *positive* and into 5 classes *very negative*, *negative*, *neutral*, *positive*, and *very positive*. Additionally, we investigated the impact of emoticons in the training and test data on system performance. For all implementations, we used Google Colab[4].

1) Mono-lingual Sentiment Analysis Systems: In the mono-lingual sentiment analysis systems, the Shona tweets were directly classified as illustrated in Figure 3. The monolingual system is based on AfroLM[5] [32] since it is the only state-of-the-art Transformer model which has been pre-trained with Shona text data.

To learn the task of sentiment analysis, we used our Shona training set for fine-tuning. We trained our AfroLM models with 4 epochs and a batch size of 32 using the AdamW optimizer [43] with an initial learning rate of 1e-5. Furthermore, we used a dropout layer for some regularisation and a fully-connected layer for our output. To get the predicted probabilities from our model, we applied a softmax function to the output.

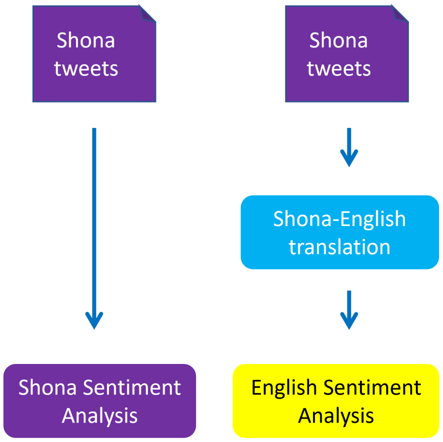

Fig. 3: Mono-lingual and Cross-Lingual Sentiment Analysis.

2) Cross-lingual Sentiment Analysis Systems: As depicted in Figure 3, in our cross-lingual systems, an English sentiment analysis system (*English Sentiment Analysis*) was the essential component, which was trained using the English translation of the labelled tweets in our training data. The tweets for testing were also machine-translated from Shona to English and then classified with the English sentiment analysis system. For the Shona-English machine translation task, we used Google's Neural Machine Translation System [44]. An overview of the system's BLEU scores over languages is given in [45]. We evaluated the quality of a subset of the English translation and obtained a BLEU score of 0.57. Reasons for the not optimal machine translation quality could be that Google's Neural Machine Translation System has difficulties with the slang and spelling mistakes that sometimes appear in the tweets. Additional human evaluations would help to improve translation accuracy. But since the human improvement of translations involves high effort and costs, our goal was to test whether we can achieve good sentiment analysis results despite non-optimal translation quality.

Our English sentiment analysis system is based on BERT[6] since it has demonstrated very good performances for mono-

[4] https://colab.research.google.com

[5] https://github.com/bonaventuredossou/MLM_AL

[6] https://huggingface.co/bert-base-uncased

System	Emoticons	Accuracy	F-score
monoLingualNoEmoticons3classes$_{AfroLM}$	no	78.60%	76.39%
monoLingualEmoticons3classes$_{AfroLM}$	yes	80.56%	79.44%
crossLingualNoEmoticons3classes$_{BERT}$	no	77.63%	76.56%
crossLingualEmoticons3classes$_{BERT}$	yes	**83.88%**	**82.34%**
monoLingualNoEmoticons5classes$_{AfroLM}$	no	68.00%	67.00%
monoLingualEmoticons5classes$_{AfroLM}$	yes	**70.38%**	**69.56%**
crossLingualNoEmoticons5classes$_{BERT}$	no	60.02%	59.23%
crossLingualEmoticons5classes$_{BERT}$	yes	61.56%	60.03%

TABLE VIII: Sentiment analysis systems for Shona.

lingual English systems and in cross-lingual systems [5] [6]. To learn the task of sentiment analysis, we used our machine-translated training set for fine-tuning. We trained our BERT models with 4 epochs and a batch size of 32 using the AdamW optimizer [43] with an initial learning rate of 2e-5. Furthermore, we used a dropout layer for some regularisation and a fully-connected layer for our output. To get the predicted probabilities from our model, we applied a softmax function to the output.

B. Results of the Sentiment Analysis for Shona

Table VIII summarises the systems' accuracies and F-scores. Our goal was to investigate the performances of the mono-lingual (*monoLingual*) and cross-lingual (*crossLingual*) systems for our 3 sentiment classes (*3classes*) and our 5 sentiment classes (*5classes*). In addition, we analysed the impact of emoticons in the tweets on the system performance by removing the emoticons in the training and test data.

The table shows that the systems where we did not remove the emoticons (*Emoticons*) outperform the same systems where we removed the emoticons (*NoEmoticons*). This shows that is important to keep the emoticons in the text data used for training and testing.

For our 3 sentiment classes (*3classes*) the cross-lingual system *crossLingualEmoticons3classes*$_{BERT}$ significantly outperforms the other systems—even though the Shona-English translation quality was not optimal with a BLEU score of 0.57. However, for our 5 sentiment classes (*5classes*) the mono-lingual system *monoLingualEmoticons5classes*$_{AfroLM}$ significantly outperforms the cross-lingual systems.

Furthermore, the performances with 3 classes are higher than with 5 classes. Our best systems achieve accuracies of 83.88% on 3 classes and 70.38% on 5 classes.

V. Conclusion and Future Work

Even though the African language Shona is spoken by nearly 17 million people, there has been neither a sentiment analysis text corpus nor sentiment analysis systems. Consequently, we collected *ShonaSenti*, our Shona sentiment analysis corpus.

In the labelling process, we experimented with distant supervision approaches based on a sentiment lexicon and emoticons. With the combination of the two processes, we were able to achieve that 23.6% of all 15,960 tweets could be correctly labelled and consequently did not have to be changed in the next step by the annotators. With these methods we could already reduce the effort of annotators and our work could be a starting point for further investigations to be even more efficient and accurate.

The following 8 topics are covered in our corpus: *sanitation*, *finance*, *agriculture*, *education*, *home affairs*, *communication*, *music* and *defense*. The *overall sentiment scores* of 8 of the 10 topics are in the negative number range. The topics *sanitation* and *education* are particularly problematic since their scores are less than -0.3, whereas with positive scores the mood regarding the topics of *home affairs* and *communication* is rather *positive*. This lays the foundation for a detailed analysis of Shona tweets, which represents the mood of the population on various topics and can be used to specifically help governmental departments to master social challenges as in [6].

Furthermore, we investigated mono-lingual and cross-lingual sentiment analysis systems for the 3 classes *negative*, *neutral* and *positive* and the 5 classes *very negative*, *negative*, *neutral*, *positive*, and *very positive*. Our best systems achieve accuracies of 83.88% on 3 classes and 70.38% on 5 classes.

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 [46]. Further our goal is to expand *SAfriSenti* with more African low-resourced languages to release a large-scale Twitter-based multilingual corpus for sentiment analysis to the NLP research community.

Ethical Impact Statement

For our data collection, individuals were asked to label the text data we collected. The participants who supported us were not dependent on the authors and participated voluntarily and free of charge. There was no conflict of interest between the supporters and the authors. For privacy reasons, the names of the supporters are not disclosed.

The collected corpus is made freely available to the community—especially to support the development of natural language processing systems for low-resourced languages. The collected text data of the corpus were not filtered out by Twitter, but it cannot be ruled out that their content is not suitable for everyone. The text data contains negative, neutral and positive sentiments in various forms. But this is the essence of a sentiment corpus that can be used to build realistic sentiment analysis systems.

References

[1] M. Wankhade, A. Rao, and C. Kulkarni, "A Survey on Sentiment Analysis Methods, Applications, and Challenges," *Artificial Intelligence Review*, pp. 1–50, 02 2022.

[2] P. P. Rokade and A. Kumari D., "Business Intelligence Analytics using Sentiment Analysis—A Survey," *International Journal of Electrical and Computer Engineering (IJECE)*, 2019.

[3] C. Zucco, H. Liang, G. D. Fatta, and M. Cannataro, "Explainable Sentiment Analysis with Applications in Medicine," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 1740–1747.

[4] J. Lalata, B. Gerardo, and R. Medina, "A Sentiment Analysis Model for Faculty Comment Evaluation Using Ensemble Machine Learning Algorithms," in *The 2019 International Conference on Big Data Engineering*, ser. BDE 2019. New York, NY, USA: ACM, 2019, p. 68–73. [Online]. Available: https://doi.org/10.1145/3341620.3341638

[5] O. Rakhmanov and T. Schlippe, "Sentiment Analysis for Hausa: Classifying Students' Comments," in *The 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022)*, Marseille, France, 2022.

[6] K. R. Mabokela and T. Schlippe, " AI for Social Good: Sentiment Analysis to Detect Social Challenges in South Africa," in *The South African Conference for Artificial Intelligence Research (SACAIR 2022)*, 12 2022.

[7] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in *The 7th Edition of the Language Resources and Evaluation Conference (LREC 2010)*, 2010, pp. 1320–1326.

[8] D. Vilares, M. A. Alonso, and C. Gómez-Rodríguez, "Sentiment Analysis on Monolingual, Multilingual and Code-switching Twitter Corpora," in *The 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2015, pp. 2–8.

[9] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 Task 4: Sentiment Analysis in Twitter," in *International Workshop on Semantic Evaluation (SemEval)*, 2016.

[10] H. Nguyen and M.-L. Nguyen, "A Deep Neural Architecture for Sentence-Level Sentiment Classification in Twitter Social Networking," in *International Conference of the Pacific Association for Computational Linguistics*, 2017, pp. 15–27.

[11] Statista, "The Most Spoken Languages Worldwide in 2022," https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide, 2022, accessed: 08-2022.

[12] K. R. Mabokela, "A Multilingual ASR of Sepedi-English Code-Switched Speech for Automatic Language Identification," in *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 2019, pp. 1–8.

[13] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource Languages: A Review of Past Work and Future Challenges," *CoRR*, vol. abs/2006.07264, 2020. [Online]. Available: https://arxiv.org/abs/2006.07264

[14] U. o. P. African Studies Center, "Shona," 2023, accessed: 01-2023. [Online]. Available: https://plc.sas.upenn.edu/shona

[15] Wikipedia, "Shona people," 2023, accessed: 01-2023. [Online]. Available: https://en.wikipedia.org/wiki/Shona_people

[16] Ethnologue, "Shona," https://www.ethnologue.com/language/sna, 2023, accessed: 01-2023.

[17] K. R. Mabokela, T. Celik, and M. Raborife, "Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape," *IEEE Access*, vol. 11, pp. 15 996–16 020, 2022.

[18] B. Mondher and O. Tomoaki, "A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter," *IEEE Access*, vol. 5, pp. 20 617 – 20 639, 2017.

[19] A. Go, R. Bhayani, and L. Huang, "Twitter Sentiment Classification using Distant Supervision," *Processing*, vol. 150, 01 2009.

[20] D. Indriani, A. H. Nasution, W. Monika, and S. Nasution, "Towards a Sentiment Analyzer for Low-Resource Languages," *CoRR*, vol. abs/2011.06382, 2020.

[21] A. Agarwal and J. S. Sabharwal, "End-to-End Sentiment Analysis of Twitter Data," in *Workshop on Information Extraction and Entity Analytics on Social Media Data*, 2012.

[22] D. Vilares, M. Alonso Pardo, and C. Gómez-Rodríguez, "Supervised Sentiment Analysis in Multilingual Environments," *Information Processing & Management*, vol. 53, 05 2017.

[23] K. R. Mabokela and T. Schlippe, "A Sentiment Corpus for South African Under-Resourced Languages in a Multilingual Context," in *The 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022)*, 06 2022, p. 70–77.

[24] A. Balahur and M. Turchi, "Comparative Experiments using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis," *Comput. Speech Lang.*, vol. 28, pp. 56–75, 2014.

[25] P. X. V. Nguyen, T. V. T. Hong, K. V. Nguyen, and N. L.-T. Nguyen, "Deep Learning versus Traditional Classifiers on Vietnamese Students' Feedback Corpus," *The 5th NAFOSTED Conference on Information and Computer Science (NICS)*, 2018.

[26] A. Kumar and A. Sharan, *Deep Learning-Based Frameworks for Aspect-Based Sentiment Analysis*. Springer Singapore, 2020, pp. 139–158.

[27] O. Rakhmanov, "A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments," *Procedia Computer Science*, vol. 178, pp. 194–204, 2020.

[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL*, 2019.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019.

[30] O. Kolchyna, T. T. P. Souza, P. C. Treleaven, and T. Aste, "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination," *arXiv: Computation and Language*, 2015.

[31] A. Kotelnikova, D. Paschenko, K. Bochenina, and E. Kotelnikov, "Lexicon-based Methods vs. BERT for Text Sentiment Analysis," in *AIST*, 2021.

[32] B. F. P. Dossou, A. L. Tonja, O. Yousuf, S. Osei, A. Oppong, I. Shode, O. O. Awoyomi, and C. Emezue, "AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages," in *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 52–64. [Online]. Available: https://aclanthology.org/2022.sustainlp-1.11

[33] Z. Lin, X. Jin, X. Xu, Y. Wang, S. Tan, and X. Cheng, "Make It Possible: Multilingual Sentiment Analysis Without Much Prior Knowledge," in *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, vol. 2, 2014, pp. 79–86.

[34] E. F. Can, A. Ezen-Can, and F. Can, "Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data," in *ACM SIGIR 2018 Workshop on Learning from Limited or Noisy Data*, 2018.

[35] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 19–27.

[36] S. K. Tasoulis, A. G. Vrahatis, S. V. Georgakopoulos, and V. P. Plagianakos, "Real Time Sentiment Change Detection of Twitter Data Streams," in *Innovations in Intelligent Systems and Applications (INISTA 2018)*, 2018, pp. 1–6.

[37] The Editors of Encyclopaedia Britannica, "Zimbabwe Summary," 2023. [Online]. Available: https://www.britannica.com/summary/Zimbabwe

[38] M. Kaity and V. Balakrishnan, "Sentiment Lexicons and non-English Languages: A Survey," *Knowledge and Information Systems*, vol. 62, 12 2020.

[39] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 168–177.

[40] L. Meetei, T. D. Singh, S. Borgohain, and S. Bandyopadhyay, "Low Resource Language Specific Pre-processing and Features for Sentiment Analysis Task, volume = 55, journal = Language Resources and Evaluation, doi = 10.1007/s10579-021-09541-9," 12 2021.

[41] S. Mohammad, "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*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 174–179.

[42] T. Perry, "LightTag: Text Annotation Platform," in *2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online and Punta Cana, Dominican Republic: ACL, 2021, pp. 20–27.

[43] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014.

[44] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," 2016. [Online]. Available: https://arxiv.org/abs/1609.08144

[45] M. W. Aiken, "An Updated Evaluation of Google Translate Accuracy," *Studies in Linguistics and Literature*, 2019.

[46] T. Schlippe, S. Ochs, and T. Schultz, "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 2012.