AI for Social Good: Sentiment Analysis to Detect Social Challenges in South Africa

Koena Ronny Mabokela[1][0000−0002−8058−969X] and Tim Schlippe[2]

[1] University of Johannesburg, South Africa
[2] IU International University of Applied Sciences, Germany
krmabokela@gmail.com, tim.schlippe@iu.org

Abstract. Sentiment analysis has the potential to help analyse people's opinions and emotions on social issues [1]. We believe that in multilingual communities sentiment analysis systems should be even used to quickly discover what social challenges exist. This would help government departments target those issues more precisely and effectively. Consequently, in this paper, we describe our experiments to apply cross-lingual sentiment analysis on South African tweets to detect social challenges described in English, Sepedi (i.e. Northern Sotho) and Setswana tweets. We investigated the polarities of the 10 most emerging topics in the tweets that fall within the jurisdictional areas of 10 South African government departments. Our AI-driven systems indicate that the topics of *employment*, *police service*, *education*, and *health* are particularly problematic for the investigated multilingual communities since more than 50% of the tweets are categorised as *negative*, whereas the mood regarding the topics of *agriculture* and *rural development* is rather *positive*. Our developed systems can be easily extended to other topics and languages.

Keywords: AI for Social Good · Sentiment Analysis · Natural Language Processing · South Africa.

1 Introduction

AI for Social Good (AI4SG) is an emerging subject of research that focuses on applying artificial intelligence (AI) to address significant social, environmental and public health issues that are facing society [2]. One can think of AI4SG as the intersection of AI with social sciences, health sciences as well as environmental sciences [3]. The field is focused on delivering positive social impact in accordance with the priorities outlined in the United Nations' Sustainable Development Goals [4] which are adopted by and applicable to South Africa. According to the 2019 Country Report, the National Development Plan has a 74% convergence with the Sustainable Development Goals and prioritises job creation, elimination of poverty, reduction of inequality and growing an inclusive economy [5]. South Africa established a national coordinating mechanism to strengthen the implementation of development policies. Therefore, South African national government departments such as *Health*, *Employment and Labour*, *Water and Sanitation*, *Rural Development*, *Education*, and *Police Service* have the mandate to support the development of these Sustainable Development Goals [5].

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* [6]. It is a growing research area of natural language processing (NLP) with a proven track record using Twitter as a source for sentiment-related text data [7–10]. Many of the sentiment analysis applications are developed for English [11]. However, English is only spoken by 19% of the world population [12] and in multilingual communities, other languages are also relevant. South Africa with its 11 official languages has a lot of multicultural and multilingual communities. However, its Niger-Congo Bantu languages are classified as low-resource languages [13] except for English and Afrikaans [14] due to the lack of digital language resources. But South African citizens also communicate and express emotions on social media platforms such as Twitter in the Bantu languages. Consequently, a sentiment analysis application that can detect sentiments from texts in South African languages to address social challenges would be extremely beneficial.

As in [15], we define a social challenge or problem as "any condition or behaviour that has negative consequences for large numbers of people and that is generally recognised as a condition or behaviour that needs to be addressed". Due to South Africa's unemployment rate of 33.9% [16], poor conditions in healthcare facilities, lack of educational tools as well as lack of water, electricity and sanitation [17, 18] and other social challenges, for this study we decided to automatically analyse the following government departments related topics with the help of AI: *employment*, *sanitation*, *police service*, *education*, *health*, *small business*, *transport*, *home affair*, *rural development*, and *agriculture*. In this paper, we present a cross-lingual sentiment analysis approach to detect the social challenges experienced in South Africa.

Our systems employ Google's Bidirectional Encoder Representations from Transformers (BERT) model [19] to classify tweets into the 3 categories *negative*, *neutral* and *positive*. BERT is a pre-trained language model which was originally trained on over 100 languages using Wikipedia text corpora. We fine-tuned BERT to perform the sentiment analysis task with the help of the *SAfriSenti* corpus [20]. *SAfriSenti* is a multilingual sentiment corpus for South African languages. For our study, we also collected a new set of tweets, over 16,000 in three languages containing our government department related topics: the *SAGovTopicTweets* corpus. Our contributions are:

- With the help of Twitter's Academic API, we collected a new corpus of tweets in 3 South African languages covering 10 South African government departments related topics—the *SAGovTopicTweets* corpus.
- We leveraged the *SAfriSenti* corpus to fine-tune the BERT model for sentiment classification.
- We utilised Sepedi-English and Setswana-English machine translation to enable cross-lingual sentiment analysis systems for Sepedi and Setswana.
- Based on the fine-tuned BERT model and the machine translation, we classified the tweet in our *SAGovTopicTweets* corpus into *negative*, *neutral* and

positive and analysed the polarity of the topics and for each topic the need to take action.
- Our results can be used as recommendations for the South African government departments to improve the social challenges identified on Twitter.

In the next section, we will describe related work. The experimental setup of our collection and sentiment analysis of tweets in English, Sepedi and Setswana will be presented in Section 3. In Section 4 we will demonstrate the results of our experiments. Finally, we will summarise our work and indicate possible future steps.

2 Related Work

AI4SG is an emerging field of study with recent successful areas such as the development of AI interventions to improve a community's well-being [21]. [22] provide a detailed analysis of approaches, use cases and examples in AI4SG. Numerous AI4SG applications use learning, reasoning, heuristic search, and problem-solving algorithms [22]. These algorithms are used by the majority of organisations and economic sectors [23]. The demand for AI applications that benefit society is great and may be used to solve several difficulties [24].

In the area of NLP for social good, [25] used sentiment analysis to automatically examine gender and race bias. [1] investigated sentiment analysis methods to classify the five major social issues of corruption, women violence, poverty, child abuse, and illiteracy. They collected English tweets and applied machine learning algorithms. The analysed text data is retrieved from microblogging services like Twitter since these sources share situational information, cover a lot of topics and contain negative, neutral and positive tweets [26, 27]. Several studies investigated different data collection methods for tweets [26, 7, 28, 8, 27]. [26] explored methods to collect millions of annotated tweets from various places, hours, and writers. Other studies used emoticons and keywords [7, 8] to extract and build Twitter-based corpora via distant supervision. [8]'s and [10]'s corpora contain code-switched tweets. To ensure correct labelling, [10] let the tweets be labelled by three annotators following the *SentiStrength* [29] strategy. [20] used the 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 [30–33]. [11] demonstrated that the Transformer models BERT [19] and RoBERTa [34] (Robustly Optimized BERT Pretraining Approach) usually outperform the other machine learning algorithms. Lexicon-based approaches were also investigated, e.g. in [35, 36], but machine learning algorithms usually perform better than the lexicon-based approaches.

Some researchers propose cross-lingual NLP approaches to solve the problems of low-resource languages by benefiting from rich-resource languages like

English [29, 30, 11, 37, 38]. 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.

In our work, we used the pre-trained BERT model [19] to build an English sentiment analysis system that is an essential component of our monolingual English and our two cross-lingual Sepedi and Setswana sentiment analysis systems.

3 Experimental Setup

In this section, we will first describe how we used the Twitter API and a word frequency-based language identification to gather South African tweets in the languages English, Sepedi and Setswana which contain the 10 most emerging topics that fall within jurisdictional areas of 10 South African government departments. Then we will present the dataset which we used to train our English sentiment analysis system which is an essential component of our monolingual and our two cross-lingual sentiment analysis systems to classify the collected tweets into *negative*, *neutral* and *positive*.

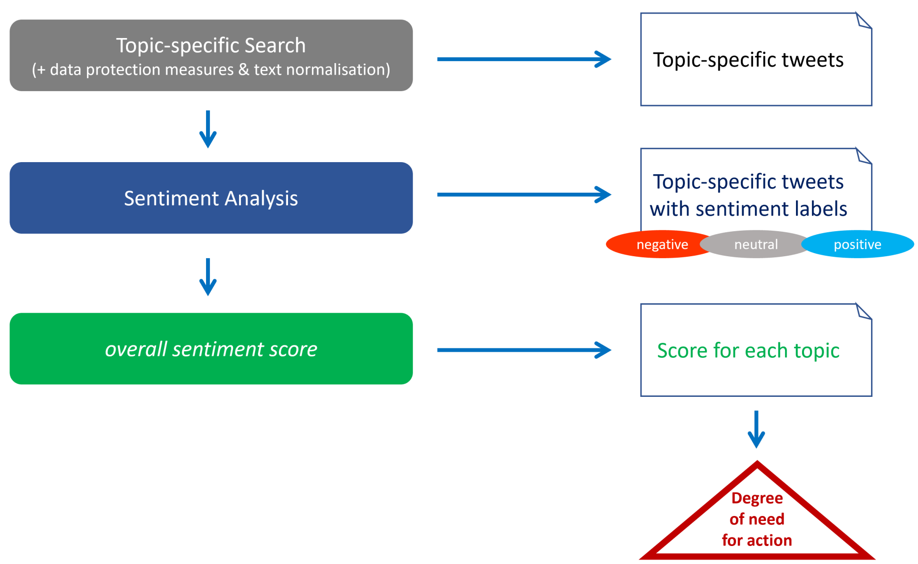

Fig. 1. Pipeline of Topic-Specific Search, Sentiment Analysis and Scoring.

3.1 Overview of Our Systems

Figure 1 shows the pipeline of our systems. In the first step, topic-specific tweets are collected with the help of search terms and data protection measures and text normalisation steps are applied. Then a sentiment analysis system classifies the tweets into *negative*, *neutral* and *positive*. In the last step, an overall sentiment score is computed for each topic that indicates the degree of need for action.

3.2 Collection of SAGovTopicTweets

For our study the goal was to collect South African tweets in English, Setswana and Sepedi covering the topics *employment*, *sanitation*, *police service*, *education*, *health*, *small business*, *transport*, *home affair*, *rural development*, and *agriculture*—the *SAGovTopicTweets* corpus. We focused on these 10 government departments related topics since they were highlighted in the State of the Nation Address for 2021 as key government issues to strengthen the economy [39].

Over the past years, Twitter has taken steps to expand and improve the Twitter API functions [40]. For example, Twitter's Academic API[3] has recently been released with the ability to get historical tweets dated back to 2006, a cap of 10 million tweets per month and more advanced filter functionality to collect relevant data [40]. The API offers the capability to collect a large pool of real-time and historic tweets for free.

Consequently, we used the Twitter API for Academic Research to crawl tweets for 24 hours which were posted between January and August 2022. To find exclusively tweets which cover our 10 government departments related topics, we searched for tweets in the following two ways:

1. We used the government departments' names[4] *Employment and Labour*, *Education*, *Police Service*, *Rural Development*, *Health*, *Small Business*, *Transport*, *Home Affairs*, *Water and Sanitation* and *Agriculture* as search terms in English and their Sepedi and Setswana translations.
2. We collected tweets which are comments on the tweets of the 10 government departments which can be found on the departments' Twitter handles.

Using Twitter's geolocation feature, we ensured that we only collected tweets from South Africa. With the help of a language identification based on word frequency [20], we also made sure that we exclusively collected tweets from our three target languages. Based on the search term that led to its download, each tweet was tagged with its topic. Once downloaded, all tweets in *SAGovTopicTweets* were subjected to our strict preprocessing steps, which we also apply in our *SAfriSenti* corpus collection. These steps are described in [20] in detail and contain data protection measures and text normalisation.

Table 1 shows three examples of government departments related South African tweets which are classified as *negative*. The first tweet is in English and covers the topic of *home affairs*. The words "very slow" and "bad" indicate a negative sentiment. The second tweet is in Sepedi and contains information on the topic of *sanitation*. The Sepedi words "re lla" (crying) and "go tura" (expensive) reflect the negative mood. The third tweet is in Setswana and belongs

[3] https://developer.twitter.com/en/use-cases/do-research/academic-research
[4] https://www.gov.za/about-government/government-system/national-departments

Table 1. Examples of government topic related South African tweets.

English tweet [*home affairs*] [*negative*]	The queue in home affairs at Wynberg is very slow one cannot get anything without waiting the service is bad
Sepedi tweet [*sanitation*] [*negative*]	Ke neng re lla ka mohlagase le makhura tsa go tura
English translation	I was crying about expensive electricity and fuel
Setswana tweet [*health*] [*negative*]	Puso ya rena ya ANC ga e re hlokomele ka tsa maphelo
English translation	Our ANC government does not take care of our health

to the topic of *health*. The Setswana word sequence "ga e re hlokomele" (not take care) describes the negative attitude.

3.3 The SAfriSenti Corpus

The *SAfriSenti* corpus is to date the largest sentiment dataset available for South African languages with 64.3% of monolingual tweets in English, Sepedi and Setswana and 36.6% of code-switched tweets between these languages [20].

For our experiments, we trained an English sentiment analysis system that is used to classify the collected English tweets plus the Sepedi and Setswana tweets translated into English. For that, we used a subset of the English part of *SAfriSenti* with 5,998 tweets that were labelled in a semi-automatic process with strict annotation guidelines [20]. To evaluate the sentiment analysis systems' performances on high-qualified labelled data which will be described in Section 3.4, we used 1,499 different English tweets, 2,106 Sepedi and 1,053 Setswana tweets from *SAfriSenti*.

3.4 Sentiment Analysis

As visualised in Figure 2, we used one monolingual and two cross-lingual sentiment analysis systems to classify the collected tweets into *negative*, *neutral* and *positive*. In all three systems, an English sentiment analysis system (*English Sentiment Analysis*) was the essential component, which was trained using our 5,998 English labelled tweets from the *SAfriSenti* corpus. In the monolingual system, the English tweets were directly classified with the English sentiment analysis system. In contrast, in the cross-lingual systems, the tweets were machine-translated from Sepedi and Setswana to English and then classified with the English sentiment analysis system. For the Sepedi-English machine translation task, we used Google's Neural Machine Translation System [41]. An overview of the system's BLEU scores over languages is given in [42]. For the Setswana-English machine translation task, we used the Autshumato Machine

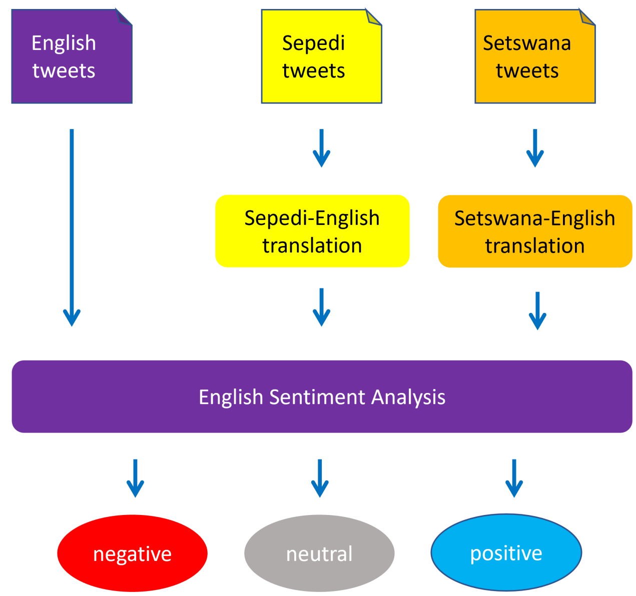

Fig. 2. Sentiment Analysis for English, Sepedi and Setswana.

Translation Web Service[5] [43] as Setswana-English machine translation was not available in Google's Neural Machine Translation System. We randomly checked the quality of the English output and can report that it was acceptable.

Our sentiment analysis system is based on the English BERT[6]. We trained our BERT models with 4 epochs and a batch size of 16 using the AdamW optimizer [44] 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. For the implementation, we used Google Colab[7].

Table 2. Performances of the sentiment analysis systems (on SAfriSenti test sets).

System	Accuracy (%)	F-score (%)
English	86.36	86.01
Sepedi	84.19	84.03
Setswana	83.26	82.74

[5] https://mt.nwu.ac.za
[6] https://huggingface.co/bert-base-uncased
[7] https://colab.research.google.com

The accuracies and F-scores of our three sentiment analysis systems on the *SAfriSenti* English, Sepedi and Setswana test sets are listed in Table 2. With accuracies between 83% and 86%, we see that the prediction of the three classes is quite acceptable with our AI-driven systems.

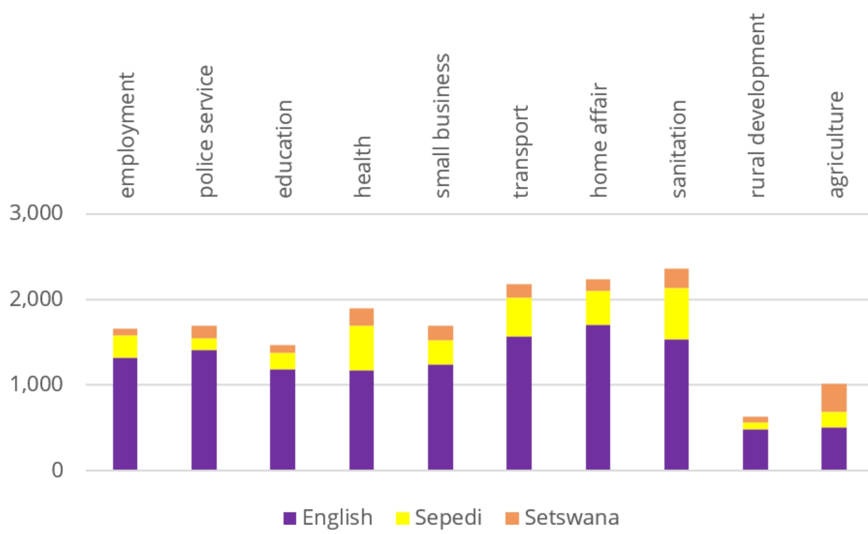

Fig. 3. Numbers of collected tweets over languages and topics.

4 Experiments and Results

In this section, we will describe how many tweets we collected to analyse the social issues. Furthermore, we will look at the benefits of classifying Sepedi and Setswana tweets in addition to English tweets. Finally, we will investigate the polarity of each investigated topic and draw conclusions about the urgency to act.

4.1 Language Distributions of Collected Tweets

Figure 3 illustrates the distribution of our collected tweets over languages and topics. In total, we collected 16,787 tweets with the Twitter API covering the topics of *employment*, *police service*, *education*, *health*, *small business*, *transport*, *home affair*, *rural development*, and *agriculture*. We see that for some topics such as *transport* (2,172 tweets), *home affair* (2,232 tweets) and *sanitation* (2,361 tweets), we found significantly more tweets, while for other topics such as *rural development* (621 tweets) and *agriculture* (1,011 tweets) much less tweets existed. Moreover, most tweets were in English (12,102 tweets, i.e. 72.1%), second most in Sepedi (3,088 tweets, i.e. 18.4%) and least in Setswana (1,597 tweets, i.e. 9.5%). We expected to see significantly more tweets in English, as many Sepedi and Setswana speakers communicate in English on the Internet. Nevertheless, we wanted to find out how strong the impact of the 27.9% tweets from Sepedi and Setswana is on the average polarities of the individual topics.

4.2 Sentiment Analysis to Detect Social Challenges

To represent the distribution of the classified tweets with only one score, we defined an *overall sentiment score* in the following way:

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

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. Based on the *overall sentiment score*, the governmental institutions can strategically decide according to which priority they will tackle the problems identified from the tweets. This lays a foundation for a recommender system that automatically analyses the polarity in text data on the internet and makes recommendations based on the value of the score where the action is needed. Of course, our formula can be extended or its factors be adapted in case of more sentiment classes (e.g., very negative, negative, neutral, positive, very positive) or if neutral tweets—which are not weights in our case—should be given more weight.

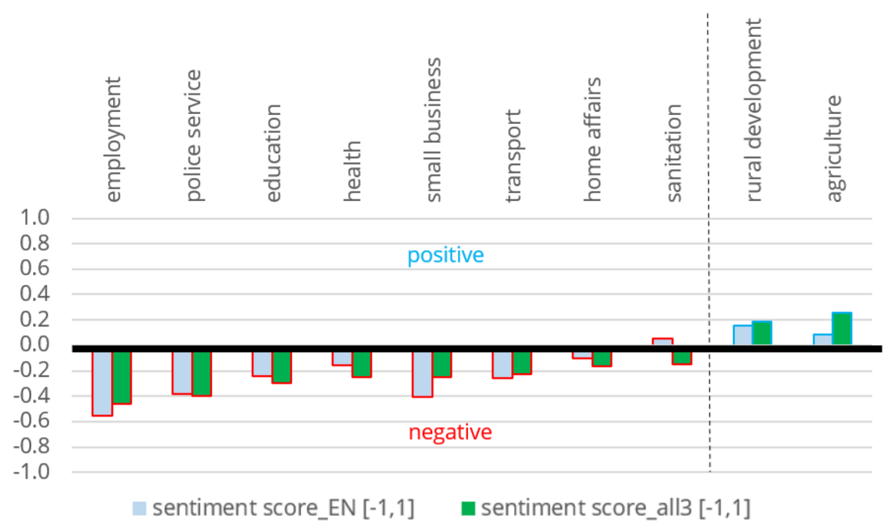

Fig. 4. Overall sentiment scores of the investigated topics.

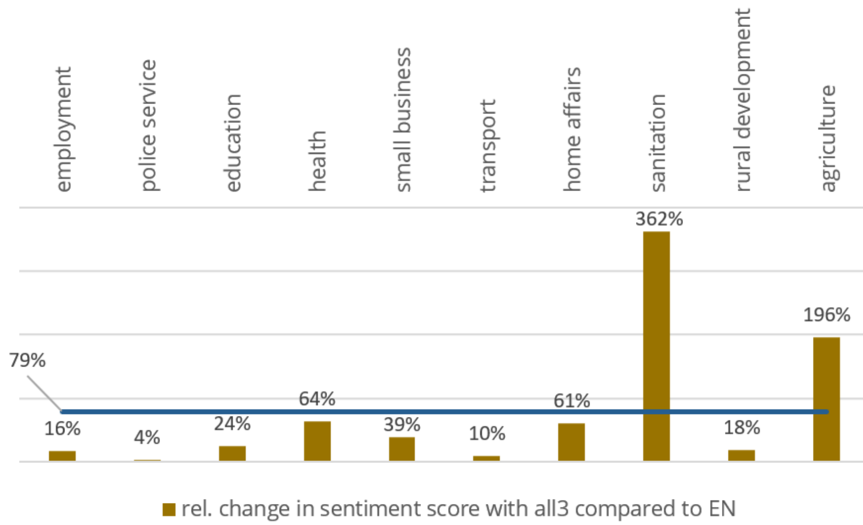

Fig. 5. Relative change in overall sentiment score with *all3* compared to *EN*.

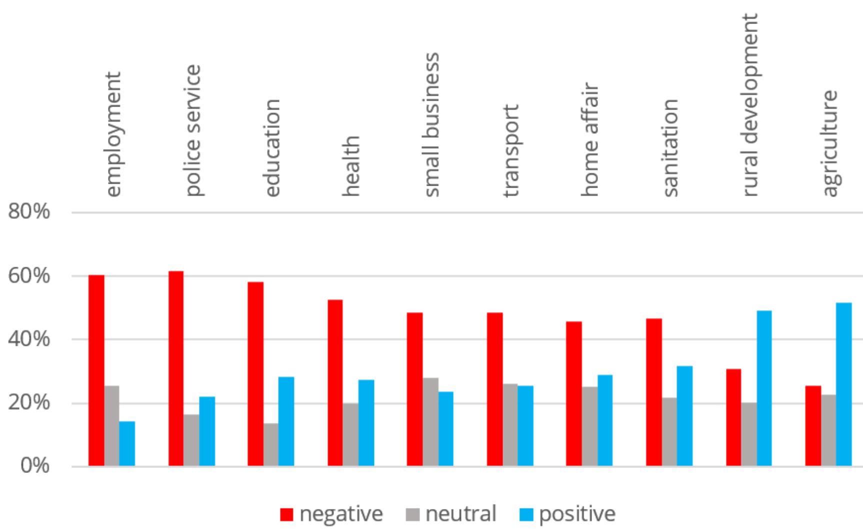

Fig. 6. Sentiment distribution of the investigated topics.

Figure 4 shows the *overall sentiment score* distribution of English tweets (*sentiment score_EN*) compared to adding Sepedi and Setswana tweets (*sentiment score_all3*) over our 10 investigated topics based on the sentiment classification of our systems. We see that the *overall sentiment scores* of 8 of the 10 topics are in the negative number range. The topics *employment* and *police service* are particularly problematic for the investigated multilingual communities since their scores are less than -0.3, whereas with positive scores the mood regarding the topics of *agriculture* and *rural development* is rather *positive*. It can also be seen that the scores differ greatly in some cases if only English tweets are considered (*EN*) and if tweets in Sepedi and Setswana are also taken into account (*all3*).

Figure 5 visualises the relative change in *overall sentiment scores* with *all3* compared to *EN*. We see that there are differences between 4% and 362%. The average deviation is 79%. This shows that it is very important to add tweets in other languages besides English to ensure a fair representative analysis.

Figure 6 shows the sentiment distribution of the classes *negative*, *neutral* and *positive* over the 10 topics analysed. Here, too, we see what the *overall sentiment scores* in Figure 4 have already indicated: The topics *employment*, *police service*, *education*, and *health* are particularly problematic for the investigated multilingual communities—more than 50% of the tweets are categorised as *negative*. The mood regarding the topics of *agriculture* and *rural development* is rather *positive*.

5 Conclusion and Future Work

With a lack of service delivery and enormous social challenges in South Africa [17], we require technologies which can assist the government to make informed decisions based on the perceptions from the citizens. Consequently, in this paper we have demonstrated that multilingual sentiment analysis is able to detect social challenges in South Africa. We investigated the polarities of the 10 most emerging topics in the tweets that fall within the jurisdictional areas of 10 South African government departments. Our AI-driven systems indicate that the topics of *employment*, *police service*, *education*, and *health* are particularly problematic for the investigated multilingual communities since more than 50% of the tweets are categorised as *negative*, whereas the mood regarding the topics of *agriculture* and *rural development* is rather *positive*.

To further understand the details of the social challenges, our next goal is to retrieve further information from the tweets with the help of information extraction algorithms. Moreover, we plan to extend our developed systems and data collections to other topics, languages and sentiment classes. Additionally, we will investigate monolingual sentiment analysis systems for Sepedi and Setswana and compare them to our cross-lingual system. While in this work we calculated the statistics about the sentiment on the topics only from the sentiment of the tweets, it can be interesting to add other features like the number of likes or the number of retweets in the analysis.

References

1. Kaur, C., Sharma, A.: Sentiment Analysis of Tweets on Social Issues using Machine Learning Approach. International Journal of Advanced Trends in Computer Science and Engineering **9**, 6303–6311 (08 2020). https://doi.org/10.30534/ijatcse/2020/310942020

2. Cowls, J., Tsamados, A., Taddeo, M., Floridi, L.: A Definition, Benchmark and Database of AI for Social Good Initiatives. Nature Machine Intelligence **3**, 111–115 (02 2021). https://doi.org/10.1038/s42256-021-00296-0
3. Tomaev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., Belgrave, D., Ezer, D., van der Haert, F.C., Mugisha, F., Abila, G., Arai, H., Almiraat, H., Proskurnia, J., Snyder, K., Otake-Matsuura, M., Othman, M.F., Glasmachers, T., de Wever, W., Teh, Y.W., Khan, M.E., Winne, R.D., Schaul, T., Clopath, C.: AI for Social Good: Unlocking the Opportunity for Positive Impact. Nature Communications **11** (2020)
4. United Nations: Sustainable Development Goals: 17 Goals to Transform our World. https://www.un.org/sustainabledevelopment/sustainabledevelopment-goals (2022), accessed: 2022-08
5. Sustainable Development Goals: Country Report 2019 – South Afrcia. Tech. Rep. ISBN 978-0-621-47619-4, Statistics South Africa (2019)
6. Wankhade, M., Rao, A., Kulkarni, C.: A Survey on Sentiment Analysis Methods, Applications, and Challenges. Artificial Intelligence Review pp. 1–50 (02 2022). https://doi.org/10.1007/s10462-022-10144-1
7. Pak, A., Paroubek, P.: Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: The 7th Edition of the Language Resources and Evaluation Conference (LREC 2010). pp. 1320–1326 (2010)
8. Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., Stoyanov, V.: SemEval-2016 Task 4: Sentiment Analysis in Twitter. In: International Workshop on Semantic Evaluation (SemEval) (2016)
9. Nguyen, H., Nguyen, M.L.: A Deep Neural Architecture for Sentence-Level Sentiment Classification in Twitter Social Networking. In: International Conference of the Pacific Association for Computational Linguistics. pp. 15–27 (2017)
10. Vilares, D., Alonso, M.A., Gómez-Rodríguez, C.: Sentiment Analysis on Monolingual, Multilingual and Code-switching Twitter Corpora. In: The 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 2–8 (2015)
11. Rakhmanov, O., Schlippe, T.: 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)
12. Statista: The Most Spoken Languages Worldwide in 2022. https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide (2022), accessed: 08-2022
13. Mabokela, K.R., Manamela, M.J.: An integrated language identification for code-switched speech using decoded-phonemes and support vector machine. In: 2013 7th Conference on Speech Technology and Human - Computer Dialogue (SpeD). pp. 1–6 (2013). https://doi.org/10.1109/SpeD.2013.6682661
14. Duvenhage, B.: Short Text Language Identification for Under Resourced Languages. In: South African Forum for Artificial Intelligence Research (FAIR 2019) (2019)
15. University of Minnesota: Social Problems. https://open.lib.umn.edu/socialproblems /chapter/1-1-what-is-a-social-problem (2022), accessed: 2022-10
16. Quarterly Labour Force Survey – Quarter 2: 2022 (2022), https://www.statssa.gov.za/publications/P0211/P02112ndQuarter2022.pdf
17. Lorraine, M., Molapo, R.: South Africa's Challenges of Realising Her Socio-Economic Rights. Mediterranean Journal of Social Sciences **5** (12 2014). https://doi.org/10.5901/mjss.2014.v5n27p900

18. Nkomo, S.: Public Service Delivery in South Africa Councillors and Citizens Critical Links in Overcoming Persistent Inequities. Tech. Rep. 42, Afrobarometer (October 2017)
19. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL (2019)
20. Mabokela, K.R., Schlippe, T.: 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). p. 70–77 (06 2022)
21. Musikanski, L., Rakova, B., Bradbury, J., Phillips, R.G., Manson, M.: Artificial Intelligence and Community Well-being: A Proposal for an Emerging Area of Research. International Journal of Community Well-Being **3**, 39–55 (2020)
22. Shi, Z.R., Wang, C., Fang, F.: Artificial Intelligence for Social Good: A Survey. CoRR **abs/2001.01818** (2020)
23. Bjola, C.: AI for Development: Implications for Theory and Practice. Oxford Development Studies **50**(1), 78–90 (2022). https://doi.org/10.1080/13600818.2021.1960960, https://doi.org/10.1080/13600818.2021.1960960
24. Hager, G., Drobnis, A.W., Fang, F., Ghani, R., Greenwald, A., Lyons, T., Parkes, D.C., Schultz, J., Saria, S., Smith, S.F., Tambe, M.: Artificial Intelligence for Social Good. ArXiv **abs/1901.05406** (2019)
25. Kiritchenko, S., Mohammad, S.M.: Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. ArXiv **abs/1805.04508** (2018)
26. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification using Distant Supervision. Processing **150** (01 2009)
27. Indriani, D., Nasution, A.H., Monika, W., Nasution, S.: Towards a Sentiment Analyzer for Low-Resource Languages. CoRR **abs/2011.06382** (2020)
28. Agarwal, A., Sabharwal, J.S.: End-to-End Sentiment Analysis of Twitter Data. In: Conference: Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data (2012)
29. Vilares, D., Alonso Pardo, M., Gómez-Rodríguez, C.: Supervised Sentiment Analysis in Multilingual Environments. Information Processing & Management **53** (05 2017). https://doi.org/10.1016/j.ipm.2017.01.004
30. Balahur, A., Turchi, M.: Comparative Experiments using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis. Comput. Speech Lang. **28**, 56–75 (2014)
31. Nguyen, P.X.V., Hong, T.V.T., Nguyen, K.V., Nguyen, N.L.T.: Deep Learning versus Traditional Classifiers on Vietnamese Students' Feedback Corpus. The 5th NAFOSTED Conference on Information and Computer Science (NICS) (2018)
32. Kumar, A., Sharan, A.: Deep Learning-Based Frameworks for Aspect-Based Sentiment Analysis, pp. 139–158. Springer Singapore (2020)
33. Rakhmanov, O.: A Comparative Study on Vectorization and Classification Techniques in Sentiment Analysis to Classify Student-Lecturer Comments. Procedia Computer Science **178**, 194–204 (2020)
34. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019)
35. Kolchyna, O., Souza, T.T.P., Treleaven, P.C., Aste, T.: Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination. arXiv: Computation and Language (2015)

36. Kotelnikova, A., Paschenko, D., Bochenina, K., Kotelnikov, E.: Lexicon-based Methods vs. BERT for Text Sentiment Analysis. In: AIST (2021)
37. Lin, Z., Jin, X., Xu, X., Wang, Y., Tan, S., Cheng, X.: 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, pp. 79–86 (2014). https://doi.org/10.1109/WI-IAT.2014.83
38. Can, E.F., Ezen-Can, A., Can, F.: Multilingual Sentiment Analysis: An RNN-Based Framework for Limited Data. In: ACM SIGIR 2018 Workshop on Learning from Limited or Noisy Data (2018)
39. Ramaphosa, C.: State of the Nation Address (2021), https://www.stateofthenation.gov.za/assets/2021/SONA%202021.pdf, accessed: 08-2022
40. Twitter Developers: A Python Wrapper around the Twitter API (2021), https://python-twitter.readthedocs.io/en/latest, accessed: 08-2022
41. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (2016), https://arxiv.org/abs/1609.08144
42. Aiken, M.W.: An Updated Evaluation of Google Translate Accuracy. Studies in Linguistics and Literature (2019)
43. Biljon, E.V., Pretorius, A., Kreutzer, J.: On Optimal Transformer Depth for Low-Resource Language Translation. In: The International Conference on Learning Representations (ICLR 2020) (2020)
44. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization (2014). https://doi.org/10.48550/ARXIV.1412.6980