Large Language Models for Sentiment Analysis to Detect Social Challenges: A Use Case with South African Languages

Koena Ronny Mabokela[1][0000−0002−8058−969X], Tim Schlippe[2][0000−0002−9462−8610], Matthias Wölfel[3][0000−0003−1601−5146]

[1] University of Johannesburg, South Africa
[2] IU International University of Applied Sciences, Germany
[3] Karlsruhe University of Applied Sciences, Germany
krmabokela@gmail.com

Abstract. Sentiment analysis can aid in understanding people's opinions and emotions on social issues. In multilingual communities sentiment analysis systems can be used to quickly identify social challenges in social media posts, enabling government departments to detect and address these issues more precisely and effectively. Recently, large-language models (LLMs) have become available to the wide public and initial analyses have shown that they exhibit magnificent *zero-shot* sentiment analysis abilities in English. However, there is no work that has investigated to leverage LLMs for sentiment analysis on social media posts in South African languages and detect social challenges. Consequently, in this work, we analyse the *zero-shot* performance of the state-of-the-art LLMs GPT-3.5, GPT-4, LlaMa 2, PaLM 2, and Dolly 2 to investigate the sentiment polarities of the 10 most emerging topics in English, Sepedi and Setswana social media posts that fall within the jurisdictional areas of 10 South African government departments. Our results demonstrate that there are big differences between the various LLMs, topics, and languages. In addition, we show that a fusion of the outcomes of different LLMs provides large gains in sentiment classification performance with sentiment classification errors below 1%. Consequently, it is now feasible to provide systems that generate reliable information about sentiment analysis to detect social challenges and draw conclusions about possible needs for actions on specific topics and within different language groups.

Keywords: AI for Social Good · Sentiment Analysis · Natural Language Processing · South Africa · Large-Language Models · LLMs.

1 Introduction

Artificial intelligence (AI) has revolutionised different areas and is now also addressing societal issues [1], with a focus on achieving the United Nations' Sustainable Development Goals (SDGs) [2]. In South Africa, the National Development Plan aligns 74% with the SDGs, emphasizing job creation, poverty elimination,

inequality reduction, and inclusive economic growth [3], with various government departments mandated to support these goals.

Sentiment analysis involves automatically detecting and classifying sentiments from textual data into categories like *negative*, *neutral*, or *positive* [4] with the help of AI and natural language processing. Applying sentiment analysis to online texts posted by citizen of a specific population can help to automatically and rapidly find and tackle social challenges in this population [5].

While sentiment analysis tools are widely available for English, which is spoken by only 19% of the global population, it is crucial to extend these tools to other languages, particularly in multilingual societies like South Africa [6]. With 12 official languages, including low-resource Niger-Congo Bantu languages [7, 8], there is a need for sentiment analysis applications that can process texts in these languages to effectively detect and address social challenges.

However, for most African languages it is very challenging to build sentiment analysis systems due to the limited availability of natural language processing corpora. Furthermore, only experts can deal with the complex algorithms required for training and fine-tuning traditional sentiment analysis systems. Given that state-of-the-art LLMs have the potential to address these problems through their growing capabilities and ease of use through prompting, particularly in *zero-shot*, we investigated their performance for sentiment analysis in African languages. Consequently, we automatically analysed the following government departments related topics with the help of state-of-the-art LLMs: *employment*, *sanitation*, *police service*, *education*, *health*, *small business*, *transport*, *home affair*, *rural development*, and *agriculture*. For each topic, we had the LLMs classify corresponding social media posts into the 3 categories *negative*, *neutral* and *positive* and compare the LLM performances to sentiment analysis systems.

For our study, we collected 16,000 social media posts from X[4], i.e. tweets, in the three languages English, Sepedi and Setswana containing our government department related topics: the *SAGovTopicTweets* corpus.

Our contributions are:

- We analyse state-of-the-art LLMs' sentiment analysis performances to detect social challenges in 3 South African languages.
- We leveraged the *SAGovTopicTweets* corpus to evaluate the performances of the tested LLMs, covering 10 South African government departments-related topics.
- Our results can be used as recommendations for the South African government departments to improve the social challenges identified on social media.

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.

[4] former Twitter

2 Related Work

AI for Social Good is a growing field of study that deals with the development of AI-based methods to enhance community well-being [9]. [10] provide a comprehensive analysis of various approaches, use cases, and examples within this field. Many AI for Social Good applications employ learning, reasoning, heuristic search, and problem-solving algorithms [10], which are widely utilised by numerous organizations and economic sectors [11]. There is a significant demand for AI applications that benefit society, as they have the potential to address numerous challenges [12].

In the field of NLP for social good, [5] utilised sentiment analysis to automatically detect gender and race bias. [13] explored sentiment analysis techniques to classify five main social issues: corruption, violence against women, poverty, child abuse, and illiteracy, collecting English tweets and applying machine learning algorithms. [14] investigated sentiment analysis for the African language Shona and reports the Shona speakers' sentiments on different topics. Text data from microblogging platforms like X (former Twitter) is often used due to its situational information, topic diversity, and range of sentiments, e.g. by [15, 16]. Various studies have examined methods for collecting tweets [15, 17–19, 16].

For automatic sentiment analysis, various machine learning algorithms such as support vector machines, decision trees, random forests, multilayer perceptrons, and long short-term memories have been examined [20–23]. [24] showed that transformer models like BERT [25] and RoBERTa [26] (Robustly Optimized BERT Pretraining Approach) generally outperform other machine learning algorithms. Lexicon-based methods, such as those investigated by [27] and [28], have also been explored, but machine learning algorithms typically yield better results than lexicon-based approaches. Some researchers suggest using cross-lingual NLP approaches to address the challenges of low-resource languages by leveraging resources from high-resource languages like English [29, 20, 24, 30, 31]. For sentiment analysis, this typically involves translating comments from the original low-resource language into English, enabling the use of well-performing models trained with extensive English resources for the classification task.

Leveraging the pre-trained English BERT model [25], in [32] we used cross-lingual sentiment analysis systems to classify tweets in English, Sepedi, and Setswana. To simplify the representation of classified tweet distributions, we defined an *overall sentiment score*, which provides a clear sentiment tendency in a single metric, facilitating topic comparisons. Government institutions can use this score to prioritise and strategically address the issues identified in the tweets. This establishes the foundation for a recommender system that automatically analyses the polarity of text data on the Internet and makes actionable recommendations based on the score. Our AI-driven systems reveal that *employment*, *police service*, *education*, and *health* are particularly problematic for the investigated multilingual communities, with over 50% of tweets categorised as *negative*, whereas topics like *agriculture* and *rural development* are seen more positively.

With the advent of LLM-based models like GPT-3.5, GPT-4, LlaMa 2, PaLM 2, and Dolly 2, there is significant potential for their use in sentiment analysis problems. [33] provide an in-depth investigation into the capabilities of LLM-based chatbots in performing various sentiment analysis tasks, ranging from conventional sentiment classification to aspect-based sentiment analysis and multifaceted analysis of subjective texts—though their study focuses solely on the English language. Their findings indicate that while LLM-based chatbots perform satisfactorily in simpler tasks, they fall short in more complex tasks requiring deeper understanding or structured sentiment information. However, LLM-based chatbots substantially outperform small language models in few-shot learning settings, highlighting their potential when annotation resources are scarce. Furthermore, [34] evaluate the English sentiment analysis performance of three state-of-the-art LLMs—GPT-3.5, GPT-4, and Llama 2—against established, high-performing transfer learning models. Their research demonstrates that, despite being *zero-shot*, LLMs can not only compete with but also, in some cases, surpass traditional transfer learning methods in sentiment classification accuracy.

The performance of Africa-centric language models against OpenAI's GPT-3.5 is evaluated by [35]. They specifically focus on their capabilities in handling low-resourced languages including the Bantu language isiZulu. The study highlights that while ChatGPT and other similar models show impressive results in high-resource languages, their performance significantly drops for African languages due to limited training data and resources. The assessment involves various tasks to demonstrate this disparity and highlights the necessity of developing more inclusive models that cater effectively to underrepresented languages. However—to the best of our knowledge—we are the first to evaluate sentiment analysis of LLMs for South African languages.

3 Experimental Setup

In this section, we will first give an overview of our system which determines the degree of action based on the sentiments of the topic-specific social media posts obtained from LLMs. Then, we will present the LLMs which we analysed for sentiment analysis. Furthermore, we will present the prompts we elaborated to instruct the LLMs to classify the social media posts. Finally, we will present the dataset which we used to evaluate the *zero-shot* performance of the analysed LLMs.

3.1 System Overview

Figure 1 illustrates the pipeline of our system. Initially, topic-specific social media posts, e.g. tweets, are gathered using search terms while ensuring data protection measures and applying text normalization steps. Following this, a sentiment analysis system categorises the tweets into *negative*, *neutral*, and *positive* sentiments. Finally, an *overall sentiment score* is calculated for each topic, indicating the degree of need for action.

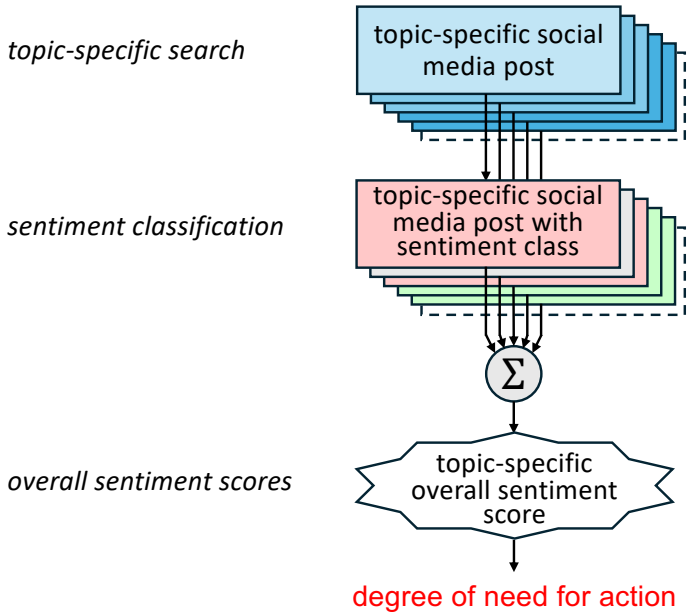

Fig. 1. Pipeline of Topic-Specific Search, Sentiment Analysis and Scoring.

To represent the distribution of the classified social media posts with only one score, we defined an

$$overall\ sentiment\ score = \frac{\#positive - \#negative}{\#allsentiments}$$

where $\#positive$ is the number of *positive* sentiments, $\#negative$ is the number of *negative* sentiments, and $\#allsentiments$ is the number of all sentiments including the number of elements classified as $\#neutral$.

The *overall sentiment score* ranges from -1 (*completely negative*) to +1 (*completely positive*), offering a clear, single metric to compare topics easily. This score helps governmental institutions prioritize which social challenges to address based on sentiment analysis from tweets. It also forms the basis for a recommender system that analyses text data polarity online and suggests actions based on the sentiment score. The formula can be adjusted to include more sentiment categories or give weight to *neutral* tweets if needed.

3.2 Large Language Models

In this subsection, we will describe the LLMs that we evaluate for sentiment analysis performance on our English, Sepedi and Setswana tweets.

GPT-3.5 GPT-3.5 was developed by OpenAI [36]. The LLM was fine-tuned using reinforcement learning from human feedback [37]. This enables the model to understand the meaning and intent behind user inquiries, resulting in relevant and useful responses. Although OpenAI has not disclosed the specific amount of training data for GPT-3.5, it is known that the prior model, GPT-3, with its 175 billion parameters, was substantially larger than other models such as BERT, RoBERTa, or T5 and was trained on 499 billion tokens [38]. The LLM can process up to 16k tokens per input [39].

GPT-4 GPT-4 is available since March 2023. It was trained on a text corpus of approximately 13 trillion tokens. This text corpus includes well-known sources like *CommonCrawl* and *RefinedWeb*, along with other undisclosed sources [40, 41]. GPT-4 was first fine-tuned using data from ScaleAI and OpenAI. Subsequently, it was fine-tuned with a reward model (Reinforcement Learning from Human Feedback) and the Proximal Policy Optimization algorithm [41, 42]. It is estimated that GPT-4 has about 1.8 trillion parameters [40, 41]. The LLM can process up to 128k tokens per input [39].

Dolly 2 The open-source LLM Dolly 2.0 was released in April 2023 [43]. Dolly is built on EleutherAI's *pythia* model series [44]. Similar to GPT-3.5, Dolly was fine-tuned to a human-created dataset [45]. The data set contains 15k manually entered entries. Through high-quality fine-tuning, Dolly 2.0 even achieves capabilities that are comparable to GPT-3.5 [45]. The LLM can process up to 2k tokens per input [45].

PaLM 2 Google's LLM PaLM 2 was trained with 1.1 trillion parameters [46] and published in May 2023 [46]. It excels in language comprehension and speech generation, demonstrating outstanding performance in both reasoning and code generation [47]. In the Bison variant which we used, the LLM can process up to 8k tokens per input [48].

LLaMa 2 The LLaMa 2 model[5], released by Meta in February 2023, features 70 billion parameters. It was fine-tuned for chat instructions using reinforcement learning from human feedback to better align with human preferences for helpfulness and safety. LLaMa 2 outperforms its predecessor, LLaMa 1, which had a maximum of 65 billion parameters [49]. Additionally, it performs exceptionally well in tests while requiring relatively little computing power [43]. The LLM can process up to 4k tokens per input [49].

Fusion of the LLM Outputs A fusion of machine learning systems' outputs has resulted in better results in other approaches, e.g. [50]. Consequently. to get a more precise sentiment classification of the social media posts, we analysed

[5] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

a fusion of the LLM outputs using majority voting. Our idea was to mitigate the misclassification by individual LLMs through this procedure. We counted the classifications for each post and selected the sentiment class chosen by the majority of LLMs as the final class.

3.3 Prompts for Sentiment Analysis

Figure 2 presents the prompts we elaborated to instruct the LLMs to classify the social media posts. For prompt engineering, it was important for us to pass a precise description to the LLMs. Consequently, for each topic, we used a separate prompt, where we added the instruction to classify each social media post, i.e. tweet in our case, indicated the topic and defined the three sentiment classes *negative*, *neutral* and *positive*. Since we detected that our analysed LLMs can handle tables in csv format well, we added the list of tweets belonging to the corresponding topic in csv format. Since our initial experiments with GPT-4 demonstrated that English prompts lead to better results than prompts in the native language—1.0% relative better F1 scores for Sepedi and 1.5% for Setswana—we decided to use English prompts. Similar findings concerning English vs. foreign prompts are reported in [51]. Note that the topic classification or the text classification, which we did manually, could also be done by the LLMs as shown in [52, 53].

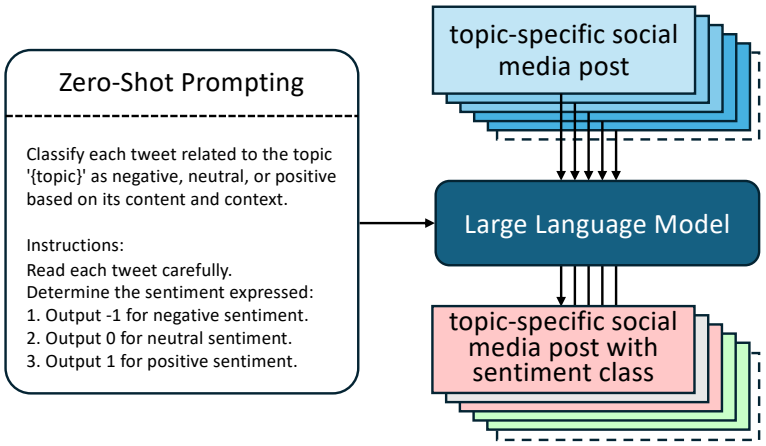

Fig. 2. *Zero-Shot* Sentiment Classification Workflow with Prompting Example and Expected Response from the LLMs.

3.4 The SAGovTopicTweets Corpus

Our goal was to analyse state-of-the-art LLMs' sentiment analysis performances to detect social challenges. Consequently, we used the *SAGovTopicTweets Cor-*

pus for our experiments which we had specifically collected for this use case as described in [32]. The *SAGovTopicTweets Corpus* contains 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*—10 government departments related topics that were highlighted in the State of the Nation Address for 2021 as key government issues to strengthen the economy [54]. Of course, English, Setswana, and Sepedi are a subset of the languages spoken in South Africa. But in the other datasets, the topics are not annotated. Nevertheless, a proof-of-concept can be achieved with this subset. The *SAGovTopicTweets Corpus* contains 16,787 tweets across languages and topics. The topics are equally distributed over the languages as described in [32]. The average number of word tokens per tweet is 21 for English, 12 for Sepedi, and 11 for Setswana, i.e. 15 on average across all three languages.

4 Experiments and Results

By employing AI for sentiment analysis, governments and other stakeholders can gain insights into the expressions of feelings and attitudes among diverse communities, which can assist in identifying and addressing social issues. Proactive analysis can facilitate timely interventions by policymakers, healthcare providers, and social workers, which can ultimately contribute to societal well-being. The accurate classification of sentiment is of paramount importance in these tasks, as it ensures a precise understanding of public emotions and reactions. Accordingly, this section presents a comprehensive technical evaluation and a socio-cultural interpretation of the sentiment analysis data. Our technical evaluation is designed to assess the methodology, accuracy, and reliability of the sentiment analysis across different languages and topics. Our socio-cultural interpretation seeks to contextualize the sentiment variations by investigating social nuances expressed in the different languages. By integrating these perspectives, our objective is to provide a comprehensive understanding of the data, emphasizing both the technological robustness and the cultural relevance of the findings.

4.1 Sentiment Classification Performances of the LLMs

To have a system which determines the degree of action based on the sentiments of the government-related topics, it is important to (1) have an excellent sentiment classification performance for all topics [55], (2) have an excellent sentiment classification performance for all languages so that all language groups are well represented [56]. Consequently, our goal was to analyze these two dimensions.

To evaluate the sentiment classification performance of the different LLMs, we conducted a study to determine their misclassification rates across different languages and topics. The results of this study are presented in Table 1. A lower value indicates less classification errors in comparison to the human-evaluated reference.

Looking at the error rates in sentiment classification, GPT-3.5 generally exhibits higher sentiment errors. GPT-4 tends to show the lowest sentiment errors across all topics (6.5%–10.9%). LLaMa 2 (9.7%–13.0%) and Dolly 2 (10.0%–12.1%) are relatively similar, often between GPT-3.5 (10.0%–13.8%) and GPT-4 (6.3%–10.9%). PaLM 2 (6.5%–12.6%) provides the second-best overall performance, right after GPT-4. According to the independent samples t-test, the overall errors in sentiment classification for Dolly 2 (11.6%) show statistical equivalence with both LLaMa 2 (11.5%) and GPT-3.5 (12.5%), indicating similar performance levels. This suggests that, despite the differences in error rates in sentiment classification per topic, the overall sentiment analysis capabilities of Dolly 2 are comparable to LLaMa 2 and GPT-3.5.

The majority voting approach in the fused system leads to lower error rates (0.2%-0.9%) for all LLMs by providing a more reliable, stable, and balanced sentiment classification, making it ideal for applications requiring consistency and robustness.

Table 1. LLMs' error rates in sentiment classification across topics and languages. Note: All annotators disagree on a subset of 1k posts in 0.6% of the tweets.

	GPT-3.5	**GPT-4**	**LLaMa 2**	**PaLM 2**	**Dolly 2**	**Fused**
agriculture	11.2%	6.5%	10.9%	8.4%	11.9%	0.3%
education	13.0%	8.4%	9.9%	8.9%	12.1%	0.5%
employment	10.6%	6.7%	10.0%	6.5%	10.3%	0.3%
health	13.5%	8.5%	11.0%	8.7%	12.5%	0.2%
home affairs	12.4%	8.6%	12.7%	10.3%	12.1%	0.4%
police service	12.9%	9.0%	12.6%	10.0%	11.0%	0.9%
rural development	13.8%	6.3%	10.5%	12.6%	11.9%	0.3%
sanitation	12.5%	7.0%	11.5%	8.9%	11.3%	0.6%
small business	12.6%	7.5%	13.0%	10.4%	11.2%	0.6%
transport	12.5%	10.9%	11.7%	8.8%	12.1%	0.6%
English	12.8%	8.6%	11.9%	9.5%	12.0%	0.4%
Sepedi	12.3%	7.0%	9.7%	8.0%	10.0%	0.7%
Setswana	10.0%	7.3%	12.2%	8.8%	11.8%	0.6%
Overall	12.5%	8.2%	11.5%	9.2%	11.6%	0.5%

Comparing the effectiveness of fusing sentiment classifications from different systems over the three languages demonstrates that the less similar the systems are, i.e. the lower their correlation, the more effective their fusion: English benefits the most (best: 8.6%, fusion: 0.4%) from the fusion of sentiment results due to the lowest correlation[6] between the classified sentiments of 0.770 among individual LLMs in comparison to 0.792 for Sepedi and 0.803 for Setswana. This low correlation indicates significant differences in the LLMs' outputs, which fusion helps to average out, leading to a more stable and reliable sentiment score.

[6] The correlations are calculated using Pearson's r.

Sepedi shows the lowest fusion gain (best: 7.0%, fusion: 0.4%) due to lower variance and higher correlation between the LLMs, meaning that individual LLMs are more equal in their outputs. Setswana has a medium level of variance and correlation, leading to moderate gains from the fusion process (best: 7.3%, fusion: 0.4%)

Even though it can be assumed that English has more training data to train the LLMs due to a much higher number of speakers (380 million native speakers) than Sepedi (4.7 million native speakers) and Setswana (6.6 million native speakers) [8], our sentiment analysis provides robust results for all three investigated languages.

Table 2. LLMs' F1-scores in sentiment classification across languages

	GPT-3.5	**GPT-4**	**LLaMa 2**	**PaLM 2**	**Dolly 2**	**Fused**
English	91.0%	94.1%	91.6%	93.2%	91.5%	97.2%
Sepedi	91.9%	95.3%	93.6%	94.5%	93.3%	97.8%
Setswana	93.2%	95.3%	92.4%	94.3%	92.5%	97.4%
Overall	91.4%	94.4%	92.0%	93.6%	91.9%	97.5%

For comparison to other studies, we have also listed the LLMs' F1 scores in Table 2. We see that the LLMs' performance is significantly better than what was reported in [32] on the SAfriSenti test sets: The BERT-based English system had an F1 score of 86.0%, the Sepedi system had an F1 score of 84.0%, and the Setswana system had an F1 score of 82.7%. This shows that the state-of-the-art LLMs can obtain better sentiment analysis performances than more traditional deep learning based approaches.

4.2 Socio-Cultural Interpretation

After demonstrating the excellent sentiment classification performance for all topics and languages—particularly with the LLMs' fusion, we conducted a socio-cultural interpretation of the sentiment distributions. Figure 3 shows *negative*, *neutral* and *positive* sentiment distributions per topic. The distributions indicate that the topics *employment*, *police service*, *education*, and *health* are particularly problematic, as more than 50% of the tweets are scored *negative*. The sentiments regarding *agriculture* and *rural development* is rather *positive*.

A better overview of the languages is visualised in the *overall sentiment scores* in Figure 4. The figure reveals significant differences in how topics are perceived across different languages. Setswana tends to have more *positive overall sentiment scores* (e.g., for *rural development* 0.30 and *agriculture* 0.64, on average the *overall sentiment score* over all topics is -0.01), while English and Sepedi exhibit a more *neutral overall sentiment score* (on average -0.18) or slightly *negative overall sentiment scores* over the topics (on average -0.29) in comparison. These variations could be influenced by cultural, socio-economic, and linguistic factors that shape how individuals express their views and opinions on different topics.

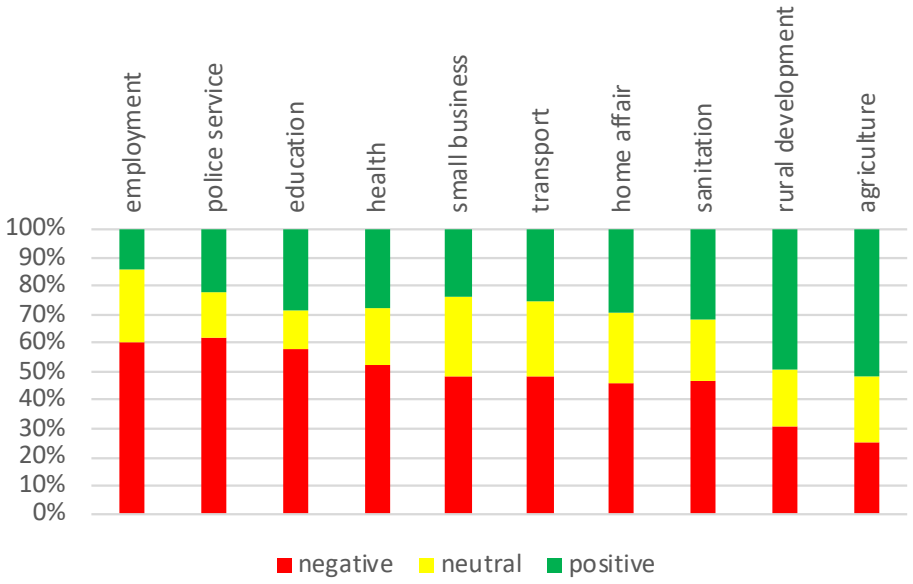

Fig. 3. Sentiment distribution of the investigated topics.

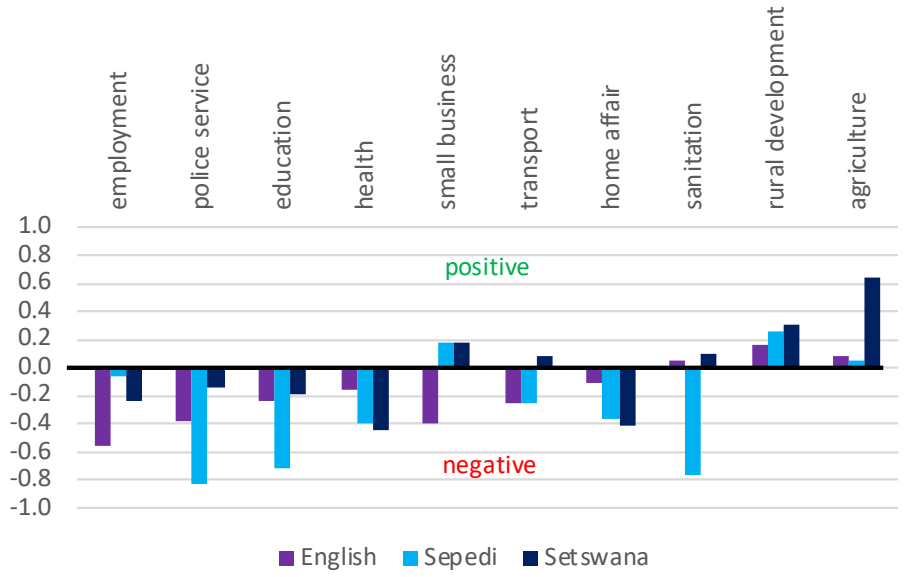

Fig. 4. *Overall sentiment score* per language.

If the recommender system identifies a *negative* sentiment towards a particular topics, the government could implement community-based initiatives to improve

service delivery specifically in the area concerned. This shows again the significance of reporting the sentiment of the individual languages and cultures.

Investigating individual topics and their relative change in *overall sentiment scores* when comparing Sepedi and Setswana to English, we observe that Sepedi provides significantly lower topic-specific *overall sentiment scores* in particular for *police service* (-0.84), *education* (-0.71), and *sanitation* (-0.77). Sepedi and English speakers may have different cultural norms and expectations regarding these services. Sepedi speakers may be more critical or have different standards for what constitutes good service. Both English and Sepedi show a move towards neutrality compared to Setswana's *positive* sentiment for *agriculture* (0.64). For *employment* English shows a particular low *overall sentiment score* (-0.55) in contrast to the other two languages.

4.3 Discussion

Our results demonstrate the usability and strong performance of the tested state-of-the-art LLMs for sentiment analysis on social media data to assess the level of need for action on social issues. For example, the F1 scores show that the LLMs' performance in sentiment analysis is comparable to other advanced systems. We would now like to briefly address some limitations that should be considered in future work to enable concrete use in a recommender system that identifies social issues.

Our analyses show that combining the outputs of multiple LLMs significantly reduces sentiment classification errors. However, this improvement in quality comes at the cost of increased computational resources, presenting a tradeoff between cost and performance. In our view, the gains in quality justify the added expense. Additionally, the analysis relies on the availability of relevant social media posts on the selected platforms, which must be accurately found and collected. To ensure realistic assessments for a given period, the posts must match the specified timeframe. Furthermore, new social media posts need to be continuously classified, although some search engines already offer features to search for specific topics. Finally, the concrete actions taken in response to the identified need for action should be defined by the responsible bodies based on the insights provided by our system.

5 Conclusion and Future Work

In conclusion, this study demonstrated the potential of state-of-the-art LLMs in addressing social challenges in South Africa by performing sentiment analysis on social media posts in English, Sepedi, and Setswana and by providing the degree of action needed in form of an *overall sentiment score*. By leveraging our SAGovTopicTweets corpus, which covers key topics related to South African government departments, the study evaluated the performance of various LLMs.

We found out that combining the outputs from our different LLMs significantly enhances sentiment classification performance, achieving sentiment classification errors below 1%. Consequently, it is now feasible to develop systems

for English, Sepedi and Setswana that reliably generate sentiment analysis information to detect social challenges and inform necessary actions across different topics and language groups using LLMs.

However, based on the diversity of the resulting *overall sentiment scores* in the topics and languages, we learned that it is important to check the sentiment for each topic and language instead of looking at them in general. The reason can be due to social environment or that different languages often express sentiments in unique ways, influenced by cultural nuances, vocabulary, and syntax.

Therefore, future work must investigate if more pronounced *negative* or *positive* sentiments between the languages for the same topics are due to linguistic and cultural differences or since the community is underserved.

References

1. Tomasev, 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)
2. United Nations: Sustainable Development Goals: 17 Goals to Transform our World. https://www.un.org/sustainabledevelopment/sustainabledevelopment-goals (2022), accessed: 2022-08
3. Sustainable Development Goals: Country Report 2019 – South Afrcia. Tech. Rep. ISBN 978-0-621-47619-4, Statistics South Africa (2019)
4. 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
5. Kiritchenko, S., Mohammad, S.M.: Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. ArXiv **abs/1805.04508** (2018)
6. Statista: The Most Spoken Languages Worldwide in 2022. https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide (2022), accessed: 08-2022
7. Mabokela, R., Roborife, M., Celik, T.: Investigating Sentiment-Bearing Words-and Emoji-based Distant Supervision Approaches for Sentiment Analysis. In: Mabuya, R., Mthobela, D., Setaka, M., Van Zaanen, M. (eds.) Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023). pp. 115–125. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.rail-1.13, https://aclanthology.org/2023.rail-1.13
8. South African Population (2022), https://census.statssa.gov.za/#/
9. 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)
10. Shi, Z.R., Wang, C., Fang, F.: Artificial Intelligence for Social Good: A Survey. CoRR **abs/2001.01818** (2020)
11. 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

12. 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)
13. 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
14. Makuwe, B., Mabokela, K.R., Schlippe, T.: Sentiment Analysis for Shona. In: 11th International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 1–8 (2023). https://doi.org/10.1109/ACII59096.2023.10388095
15. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification using Distant Supervision. Processing **150** (01 2009)
16. Indriani, D., Nasution, A.H., Monika, W., Nasution, S.: Towards a Sentiment Analyzer for Low-Resource Languages. CoRR **abs/2011.06382** (2020)
17. 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)
18. 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)
19. 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)
20. Balahur, A., Turchi, M.: Comparative Experiments using Supervised Learning and Machine Translation for Multilingual Sentiment Analysis. Comput. Speech Lang. **28**, 56–75 (2014)
21. 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)
22. Kumar, A., Sharan, A.: Deep Learning-Based Frameworks for Aspect-Based Sentiment Analysis, pp. 139–158. Springer Singapore (2020)
23. 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)
24. 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)
25. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL (2019)
26. 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)
27. 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)
28. Kotelnikova, A., Paschenko, D., Bochenina, K., Kotelnikov, E.: Lexicon-based Methods vs. BERT for Text Sentiment Analysis. In: AIST (2021)
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. 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
31. 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)
32. Mabokela, K.R., Schlippe, T.: AI for Social Good: Sentiment Analysis to Detect Social Challenges in South Africa. In: Pillay, A., Jembere, E., Gerber, A. (eds.) Artificial Intelligence Research. pp. 309–322. Springer Nature Switzerland, Cham (2022)
33. Zhang, W., Deng, Y., Liu, B., Pan, S., Bing, L.: Sentiment Analysis in the Era of Large Language Models: A Reality Check. In: Duh, K., Gomez, H., Bethard, S. (eds.) Findings of the Association for Computational Linguistics: NAACL 2024. pp. 3881–3906. Association for Computational Linguistics, Mexico City, Mexico (Jun 2024), https://aclanthology.org/2024.findings-naacl.246
34. Krugmann, J., Hartmann, J.: Sentiment analysis in the age of generative ai. Customer Needs and Solutions **11**(3) (2024). https://doi.org/10.1007/s40547-024-00143-4, https://doi.org/10.1007/s40547-024-00143-4
35. Abbott, J., Dossou, B., Mbuya, R.: Comparing Africa-centric Models to OpenAI's GPT-3.5. https://lelapa.ai/comparing-africa-centric-models-to-openais-gpt3-5-2 (2023), lelapa AI, Accessed: 2024-07-29
36. Baidoo-Anu, D., Owusu Ansah, L.: Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. SSRN 4337484 (2023)
37. OpenAI: What is ChatGPT? (2023), https://help.openai.com/en/articles/6783457-what-is-chatgpt
38. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners. CoRR **abs/2005.14165** (2020)
39. Nwanne, W.: Comparing gpt-3.5 & gpt-4: A thought framework on when to use. AI Azure AI Services Blog (2023), https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/comparing-gpt-3-5-amp-gpt-4-a-thought-framework-on-when-to-use/ba-p/4088645
40. Patel, D., Wong, G.: GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE. https://github.com/llv22/gpt4_essay/blob/master/GPT-4-4.JPG (July 2023), accessed: 30-09-2023
41. Yalalov, D., Myakin, D.: GPT-4's Leaked Details Shed Light on its Massive Scale and Impressive Architecture. Metaverse Post (July 2023), https://mpost.io/gpt-4s-leaked-details-shed-light-on-its-massive-scale-and-impressive-architecture/#gpt-4s-massive-parameters-count
42. OpenAI: GPT-4 (March 2023), https://openai.com/research/gpt-4
43. Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.Y., Wen, J.R.: A Survey of Large Language Models (2023)

44. Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M.A., Purohit, S., Prashanth, U.S., Raff, E., Skowron, A., Sutawika, L., van der Wal, O.: Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In: The 40th International Conference on Machine Learning. Honolulu, Hawaii, USA (2023)
45. Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., Xin, R.: Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM (2023), https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm
46. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J.H., Shafey, L.E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., et al.: PaLM 2 Technical Report (2023)
47. Narang, S., Chowdhery, A.: Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance (April 2022), https://blog.research.google/2022/04/pathways-language-model-palm-scaling-to.html
48. Google: PaLM Documentation (2024), https://ai.google.dev/palm_docs/palm, accessed: 2024-07-29
49. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models (2023)
50. Agyemang, A., Schlippe, T.: AI in Education: An Analysis of Large Language Models for Twi Automatic Short Answer Grading. In: Artificial Intelligence Research. Springer Nature Switzerland, Cham (2024)
51. Kmainasi, M.B., Khan, R., Shahroor, A.E., Bendou, B., Hasanain, M., Alam, F.: Native vs Non-Native Language Prompting: A Comparative Analysis (2024), https://arxiv.org/abs/2409.07054
52. Fatemi, B., Rabbi, F., Opdahl, A.L.: Evaluating the Effectiveness of GPT Large Language Model for News Classification in the IPTC News Ontology. IEEE Access **11**, 145386–145394 (2023). https://doi.org/10.1109/ACCESS.2023.3345414
53. Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., Wang, G.: Text Classification via Large Language Models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 8990–9005. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.603, https://aclanthology.org/2023.findings-emnlp.603
54. Ramaphosa, C.: State of the Nation Address (2021), https://www.stateofthenation.gov.za/assets/2021/SONA%202021.pdf, accessed: 08-2022
55. Bhatia, S., P, D.: Topic-specific sentiment analysis can help identify political ideology. In: Balahur, A., Mohammad, S.M., Hoste, V., Klinger, R. (eds.) Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 79–84. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-6212, https://aclanthology.org/W18-6212
56. Vilares, D., Alonso, M.A., Gómez-Rodríguez, C.: Supervised sentiment analysis in multilingual environments. Information Processing & Management **53**(3), 595–607 (2017). https://doi.org/https://doi.org/10.1016/j.ipm.2017.01.004, https://www.sciencedirect.com/science/article/pii/S0306457316302540