

Language-Independent Sentiment Labelling with Distant Supervision: A Case Study for English, Sepedi and Setswana

Koena Ronny Mabokela*, Tim Schlippe[§], Mpho Raborife*, Turgay Celik^{||}

*University of Johannesburg, South Africa

[§]IU International University of Applied Sciences, Germany

^{||}University of the Witwatersrand, South Africa
krmabokela@gmail.com

Abstract

Sentiment analysis is a helpful task to automatically analyse opinions and emotions on various topics in areas such as *AI for Social Good*, *AI in Education* or marketing. While many of the sentiment analysis systems are developed for English, many African languages are classified as low-resource languages due to the lack of digital language resources like text labelled with corresponding sentiment classes. One reason for that is that manually labelling text data is time-consuming and expensive. Consequently, automatic and rapid processes are needed to reduce the manual effort as much as possible making the labelling process as efficient as possible. In this paper, we present and analyze an automatic language-independent sentiment labelling method that leverages information from sentiment-bearing emojis and words. Our experiments are conducted with tweets in the languages English, Sepedi and Setswana from *SAfriSenti*, a multilingual sentiment corpus for South African languages. We show that our sentiment labelling approach is able to label the English tweets with an accuracy of 66%, the Sepedi tweets with 69%, and the Setswana tweets with 63%, so that on average only 34% of the automatically generated labels remain to be corrected.

Keywords: sentiment analysis, artificial intelligence, natural language processing, South African languages

1. Introduction

Sentiment analysis helps analyze and extract information about polarity from textual feedback and opinions. Sentiment analysis draws attention in business environments (Rokade and Kumari, 2019) and other areas, like medicine (Zucco et al., 2018), education (Mabokela et al., 2022; Rakhmanov and Schlippe, 2022) and AI for Social Good (Mabokela and Schlippe, 2022b).

Sentiment analysis for under-resourced language still is a skewed research area. Although, there are some considerable efforts in emerging African countries to develop resources for under-resourced languages, some languages such as indigenous South African languages still suffer from a lack of datasets. One reason for that is that manually labelling text data is time-consuming and expensive. Consequently, automatic and rapid processes are needed to reduce the manual effort as much as possible making the labelling process as efficient as possible. In this paper, we present and analyze an automatic language-independent sentiment labelling algorithm that leverages information from sentiment-bearing emojis¹ and words. We will evaluate our algorithm on a subset of our *SAfriSenti* corpus (Mabokela and Schlippe, 2022a; Mabokela and Schlippe, 2022b) with English, Sepedi and Setswana tweets. Sepedi is mainly spoken in the northern parts of South Africa by 4.7 million people and Setswana by 4.5 million people (Statista, 2022).

In the next section, we will describe related work. In section 3 we will present our language-independent algorithm

for sentiment labelling. The experimental setup will be characterised in Section 4. In Section 5 we will summarise the results of our experiments. We will conclude our work in Section 6 and indicate possible future work.

2. Related Work

Previous studies investigated sentiment data collection strategies for under-resourced languages on Twitter (Pak and Paroubek, 2010; Vosoughi et al., 2016). The methods focus on labelling only two sentiment classes —positive and negative. Meanwhile other research work has explored strategies to label three sentiment classes in Twitter—positive, neutral, and negative —using human annotators (Pang et al., 2002; Pak and Paroubek, 2010; Vilares et al., 2016; Nakov et al., 2019). Despite the attempt to automate the data labelling process (Kranjc et al., 2015), the hand-crafted annotation is to date the most preferred method of data labelling in many natural language processing tasks (Chakravarthi et al., 2020). However, manual annotation presents challenges and it is deemed an expensive process. Notably, (Jamatia et al., 2020; Gupta et al., 2021) employed manually annotated tweets, while other studies focus on automated data labelling solutions (Kranjc et al., 2015). (Vosoughi et al., 2016) investigated various pipelines to collect data on Twitter using distant supervised learning. In this approach, they use positive and negative emojis as indicators to annotate tweets.

(Go et al., 2009) explored distant supervision methods to label millions of tweets using positive and negative search terms (i.e. term queries) in the Twitter API and emojis to pre-classify the tweets. (Vilares et al., 2016) investigated *SentiStrength* scores to label an English-Spanish code-switching Twitter corpus. *SentiStrength* is an online

¹Emojis are pictorial representations of emotions, ideas, or objects in electronic communication to add emotional context.

sentiment analysis system available for a few languages (Thelwall et al., 2011).

Compared to (Vilares et al., 2016; Cliche, 2017; Jamatia et al., 2020), we also investigate a distant supervised annotation method. However, we automatically build up lists with sentiment-bearing words after we have made use of emojis as indicators for the sentiment classes. This way we can label all tweets—first those tweets that contain sentiment-bearing emojis, and then the rest of the tweets based on the words in the tweets with the sentiment-bearing emojis.

3. Sentiment Labelling with Distant Supervision

As illustrated in Figure 1, 2 and 3, we propose the following algorithm for sentiment labelling that leverages information from sentiment-bearing emojis and words:

- *Step 1_{emojis}*: Classify tweets with sentiment-bearing emojis into the classes *negative*, *neutral* and *positive* (Figure 1).
- *Step 2_{lists}*: Create lists with sentiment-bearing words (Figure 2):
 1. Collect all words from *negative*, *neutral* and *positive*.
 2. Then remove words that occur in one or both other lists.
- *Step 3_{words}*: Classify remaining tweets without sentiment-bearing emojis into the classes *negative*, *neutral* and *positive* based on the highest word coverage with the lists of sentiment-bearing words (Figure 3).

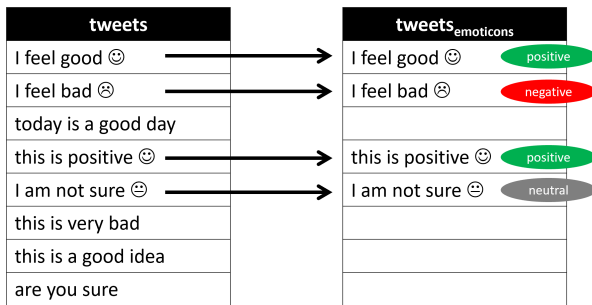


Figure 1: Classify tweets with sentiment-bearing emojis into the 3 classes (*step1_{emojis}*).

4. Experimental Setup

In this section, we will describe the dataset for our experiments and how we used emojis as indicators for sentiments.

4.1. SAfriSenti

To evaluate our algorithm for sentiment labelling with comparable numbers of tweets in three languages, we applied it to monolingual 7,000 English tweets, 7,000 Sepedi tweets and 7,000 Setswana tweets from the *SAfriSenti* corpus. *SAfriSenti* is to date the largest sentiment dataset available

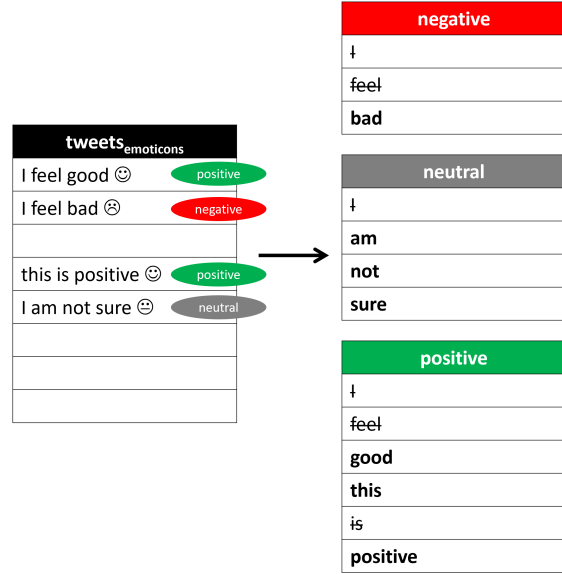


Figure 2: Create lists with sentiment-bearing words (*step2_{lists}*).

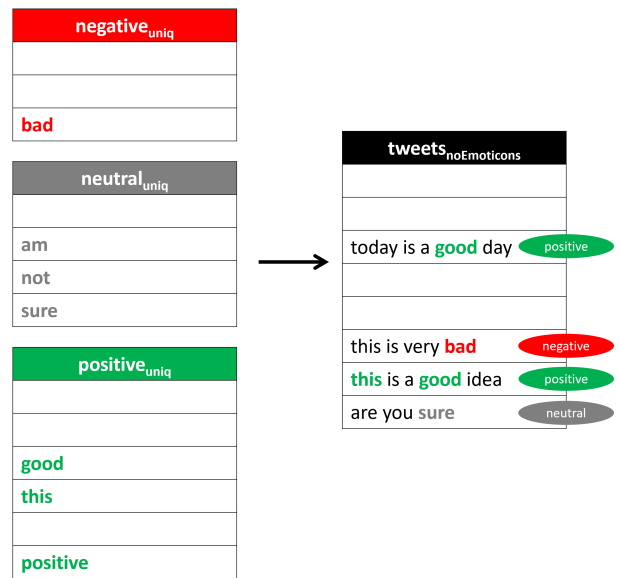


Figure 3: Sentiment-bearing words as indicators for remaining tweets' sentiment classes (*step3_{words}*).

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 (Mabokela and Schlippe, 2022a). The monolingual tweets' distributions of the classes *negative*, *neutral*, *positive* are demonstrated in Table 1, 2 and 3.

Class	Number	%
positive	2,052	29.3
negative	3,448	49.3
neutral	1,500	21.4
Total	7,000	

Table 1: Distribution of English tweets.

Class	Number	%
positive	3,500	50.0
negative	2,270	32.4
neutral	1,230	17.6
Total	7,000	

Table 2: Distribution of Sepedi tweets.

Class	Number	%
positive	3,230	46.1
negative	2,180	31.1
neutral	1,590	22.8
Total	7,000	

Table 3: Distribution of Setswana tweets.

4.2. Emojis

For our experiments, we defined 12 emojis as *negative* indicators, 10 emojis as *neutral* indicators, and 12 emojis as *positive* indicators as listed in Table 4, for which we were sure that they would represent the corresponding sentiments well. Of course, our emoji list can be extended based on further information such as the Emoji Sentiment Ranking² (Kralj Novak et al., 2015).

sentiment	#emojis
negative	12
neutral	10
positive	12

Table 4: Emojis for *negative*, *neutral* and *positive*.

A subset of such emojis is illustrated in Figure 4. To be platform-independent, our algorithm finds and compares the emojis in Unicode. If a tweet contains multiple emojis representing different sentiments, the tweet is labelled with the sentiment class which has the most emojis in the tweet. The tweet is not labelled at this step if there is no majority.

Class	Emoticons
Negative	👎 🙄 🙃 🙇 🙈
Neutral	😐 😏 😬 😇 😊 😄
Positive	😁 😂 🥰 🥳 🥰 🥳

Figure 4: Examples of emojis

5. Experiments and Results

We evaluated our algorithm for sentiment labelling on the English, Sepedi and Setswana tweets from *SAfriSenti*. Table 5 demonstrates the absolute numbers and percentages of tweets with sentiment-bearing emojis in *step 1_emojis* and the absolute numbers and percentages of the remaining tweets which were labelled using our sentiment-bearing words in *step 3_words*.

²https://kt.ijs.si/data/Emoji_sentiment_ranking

language	<i>step 1_emojis</i>	<i>step 3_words</i>
English	4,210 (60.1%)	2,790 (39.9%)
Sepedi	5,871 (83.9%)	1,129 (16.1%)
Setswana	3,249 (46.4%)	3,751 (53.6%)

Table 5: #tweets and %tweets for *step1* and *step3*.

language	<i>step 1_emojis</i>	<i>step 3_words</i>	<i>step 1-to-3</i>
English	68.7%	63.5%	66.2%
Sepedi	69.5%	64.6%	68.7%
Setswana	66.1%	59.7%	62.7%

Table 6: Accuracies.

language	<i>step 1_emojis</i>	<i>step 3_words</i>	<i>step 1-to-3</i>
English	66.8%	62.2%	64.6%
Sepedi	68.2%	63.4%	67.9%
Setswana	65.2%	58.7%	61.5%

Table 7: F-scores.

Table 6 and Table 7 shows the accuracies and F-scores of automatically classifying tweets with sentiment-bearing emojis in *step 1_emojis*, the accuracies and F-scores of automatically classifying the remaining tweets using our sentiment-bearing words in *step 3_words* as well as the accuracies and F-scores of all 7k labeled tweets together (*step 1-to-3*). For computing the F-scores, we took the *Macro F₁ Score Calculation*, i.e. the average of each class’s *F₁* score.

The results of the two tables demonstrate that the quality of the automatic labeling in *step 1_emojis* is better than that in *step 3_words*. This shows that in *step 2_lists* the automatically created word lists are worse indicators than the emojis. The goal of our algorithm is to label a large part correctly so that the annotators no longer have to label all labels from scratch and thus minimize the workload. From the accuracies and the F-scores, we see that we were able to reduce the manual effort, but there is still room for improvement: With an accuracy of 66%, the annotators would still have to change 34% of the labelled tweets for English. Our algorithm performs best for Sepedi with 68% accuracy, which would require 32% changes. With Setswana, the accuracy is 63%, which is why 37% of the labels would have to be changed.

Both the quality of *step 1_emojis* but also the quantity of tweets with sentiment-bearing emojis in the corpus, based on which the word lists are generated in *step 2_lists*, have an impact on the quality of the labeling in *step 3_words*: The more tweets with sentiment-bearing emojis, the higher the chance that qualified sentiment-bearing words remain in the lists after *step 2_lists*. We believe that Sepedi especially performs best since out of the 7k Sepedi tweets, 84% contain sentiment-bearing emojis and thus more text can be used for the generation of the word lists than for English (60%) and Setswana (47%). Still, with an accuracy of 70%, the challenge remains to figure out how to use emojis as a better indicator.

Adding more features for pre-labeling such as a translation

of the tweets or word lists translated from other languages could help. But then the algorithm would no longer be completely language-independent and would have to deal with information coming from outside the corpus.

6. Conclusion and Future Work

In this paper, we have presented a language-independent algorithm for sentiment labelling. Our algorithm uses sentiment-bearing emojis as initial features to build lists with sentiment-bearing words. Since our approach is only based on frequencies, no training of a machine learning system is required. This way we completely avoid any manual or higher computational effort.

Our analyses on the under-resourced languages Sepedi and Setswana plus English demonstrated that accuracies between 63% and 69% are possible using our distant supervision approach. This significantly reduces the manual effort to label tweets with sentiment classes since the human annotators need to change between 31%–37% of the tweets that have been pre-labelled with our algorithm instead of adding all labels from scratch. After our proof of concept with three languages, it can be assumed that our approach works for other languages as well, since people often use emojis in their posts—no matter what language their posts are.

Consequently, it is interesting to apply our approach to more languages and to experiment with cross-lingual features. Future work may also include investigating whether it is helpful to label a tweet as *neutral* if it contains a comparable number of *positive* and *negative* emojis. Our current algorithm does not assign a label to the tweet in this case. Furthermore, we plan to combine our approach with active learning, i.e. an iterative process where annotators manually classify tweets which are then used to re-train machine learning systems for classification. Additionally, our goal is to create a multilingual natural language processing model and investigate the synergy effects across languages in sentiment analysis. (Makgatho et al., 2021)’s word embeddings could be a good basis for such a model.

References

- Chakravarthi, Bharathi Raja, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P. McCrae, 2020. Corpus creation for sentiment analysis in code-mixed tamil-english text. In *SLTU*.
- Cliche, Mathieu, 2017. BB.twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs. In *SemEval*.
- Go, Alec, Richa Bhayani, and Lei Huang, 2009. Twitter Sentiment Classification Using Distant Supervision. *Processing*, 150.
- Gupta, Akshat, Sargam Menghani, Sai Krishna Rallabandi, and Alan W. Black, 2021. Unconscious Self-Training for Sentiment Analysis of Code-Linked Data. *ArXiv*, abs/2103.14797.
- Jamatia, Anupam, Steve Durairaj Swamy, Bjorn Gamback, Amitava Das, and Swapan Debbarma, 2020. Deep Learning Based Sentiment Analysis in a Code-Mixed English-Hindi and English-Bengali Social Media Corpus. *International Journal on Artificial Intelligence Tools*, 29.
- Kralj Novak, Petra, Jasmina Smailović, Borut Sluban, and Igor Mozetič, 2015. Sentiment of Emojis. *PLoS ONE*, 10(12).
- Kranjc, Janez, Jasmina Smailovic, Vid Podpecan, Miha Grcar, Martin Znidari, and Nada Lavrac, 2015. Active Learning for Sentiment Analysis on Data Streams: Methodology and Workflow Implementation in the CloudFlows Platform. *Inf. Process. Manag.*, 51:187–203.
- Mabokela, Koena Ronny, Turgay Celik, and Mpho Raborife, 2022. Multilingual Sentiment Analysis for Under-Resourced Languages: A Systematic Review of the Landscape. *IEEE Access*:1–22.
- Mabokela, Koena Ronny and Tim Schlippe, 2022a. A Sentiment Corpus for South African Under-Resourced Languages in a Multilingual Context. In *SIGUL 2022*.
- Mabokela, Koena Ronny and Tim Schlippe, 2022b. AI for Social Good: Sentiment Analysis to Detect Social Challenges in South Africa. In *SACAIR*.
- Makgatho, Mack, Vukosi Marivate, Tshephisho Sefara, and Valencia Wagner, 2021. Training Cross-Lingual embeddings for Setswana and Sepedi. *DHASA*, 3(03).
- Nakov, Preslav, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov, 2019. SemEval-2016 task 4: Sentiment analysis in Twitter. *arXiv preprint arXiv:1912.01973*.
- Pak, Alexander and Patrick Paroubek, 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREC*.
- Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan, 2002. Thumbs up? Sentiment Classification Using Machine Learning Techniques. In *EMNLP*.
- Rakhmanov, Ochilbek and Tim Schlippe, 2022. Sentiment Analysis for Hausa: Classifying Students’ Comments. In *SIGUL 2022*. Marseille, France.
- Rokade, Prakash P. and Aruna Kumari, 2019. Business Intelligence Analytics Using Sentiment Analysis—A Survey. *IJECE*.
- Statista, 2022. African Countries with the Largest Population as of 2020.
- Thelwall, Mike A, Kevan Buckley, and Georgios Paltoglou, 2011. Sentiment in Twitter Events. *J. Assoc. Inf. Sci. Technol.*, 62:406–418.
- Vilares, David, Miguel A Alonso, and Carlos Gómez-Rodríguez, 2016. EN-ES-CS: An English-Spanish Code-Switching Twitter Corpus for Multilingual Sentiment Analysis. In *Tenth International Conference on Language Resources and Evaluation (LREC’16)*.
- Vosoughi, Soroush, Helen Zhou, and Deb Roy, 2016. Enhanced Twitter Sentiment Classification Using Contextual Information. *CoRR*, abs/1605.05195.
- Zucco, Chiara, Huizhi Liang, Giuseppe Di Fatta, and Mario Cannataro, 2018. Explainable Sentiment Analysis with Applications in Medicine. In *BIBM*.