Exploring Reproducibility of Human-Labelled Data for Code-Mixed Sentiment Analysis

Sachin Sasidharan Nair, Tanvi Dinkar, Gavin Abercrombie
Heriot-Watt University, Edinburgh, Scotland
{ss2246, t.dinkar, g.abercrombie}@hw.ac.uk

Abstract

Growing awareness of a ‘Reproducibility Crisis’ in natural language processing (NLP) has focused on human evaluations of generative systems. While labelling for supervised classification tasks makes up a large part of human input to systems, the reproduction of such efforts has thus far not been been explored. In this paper, we re-implement a human data collection study for sentiment analysis of code-mixed Malayalam movie reviews, as well as automated classification experiments. We find that missing and under-specified information makes reproduction challenging, and we observe potentially consequential differences between the original labels and those we collect. Classification results indicate that the reliability of the labels is important for stable performance.

Keywords: Reproducibility, Human Data Collection, Sentiment Analysis, Malayalam

1. Introduction

There has recently been growing awareness of a ‘Reproducibility Crisis’ in natural language processing (NLP) (Belz et al., 2021). This has focused on the apparent impossibility of reproducing human evaluation studies of the outputs of natural language generation (NLG) systems (Belz et al., 2023; Thomson et al., 2024). However, while text labelling makes up the largest part of human input to NLP projects, there have been almost no reported attempts (to our knowledge) to reproduce human label collection for NLP tasks outwith NLG evaluation.

In this study, part of ReproNLP[1] Track A (Belz and Thomson, 2024), we focus on one of the most active areas of NLP over the last two decades, sentiment analysis (Mäntylä et al., 2018): an NLP task that aims to categorise the sentiment expressed in textual data (Liu, 2012).

In addition, we delve into the complexities introduced to sentiment analysis by using code-mixed language data. We re-examine the Malayalam-English corpus of Chakravarthi et al. (2020), classified with one of five distinct labels (*Positive*, *Negative*, *Neutral*, *Mixed feelings*, and *Non-Malayalam*). We assess the challenges faced while re-annotating the original corpus, and while reproducing the processes followed by the original study including the annotation process and the re-implementation of automated classifiers, verifying whether we are able to achieve similar results to those of the original study.

An example item from the corpus is shown here:

> Ufff vere level ikkaaa ingha pwoli aahn
> *Another level, ikka you are awesome*
> Assigned Label: *Positive*

2. Background & Related Work

2.1. Reproducbility in NLP

In response to the widespread reproduciblity issues uncovered in other scientific fields (Baker, 2016), there have been increasing efforts to establish standards for reproducibility in NLP, such as workshops that aim to tackle these problems (e.g. and Machine Learning Reproducbility Challenge (MLRC), HumEval)(Belz et al., 2021). Other initiatives, such as reproducibility checklists[2] have been adopted at major conferences such as EMNLP and AAAI to foster the integrity and validity of experiments.

The conversation around reproducibility is nuanced. In their review, Belz et al. (2021) note that the definitions provided by six different sources had varied interpretations of reproducibility and replicability, lacking standardised definitions. This diversity further complicates the efforts to establish consistent reproduciblity practices in the field of NLP. Moreover, the discussions of Rougier et al. (2017) and Wieling et al. (2018) highlight the need for a common understanding that also involves transparency and openness to guide reproducibility efforts.

Gundersen and Kjensmo (2018) evaluated 400 research papers from major conferences IJCAI and AAAI revealing a lack of comprehensive documentation. Only an average of 20% to 30% of necessary variables were documented, which indicated a significant gap. Although there was a slight improvement in documentation over time, the reproducibility scores generally decreased as documentation requirements grew. This analysis confirms that lack of documentation is a significant

[1] https://repronlp.github.io

[2] https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

Proceedings of the HumEval @LREC-COLING 2024, pages 114–124
Turin, Italy. May 21st, 2024. ©2024 European Language Resources Association (ELRA)

challenge faced in AI research reproducibility.

As this field progresses, it is clear that we must focus on addressing the challenges by the use of continued dialogue and action to develop standards for reproducibility. In this paper, we begin to shift the focus of NLP reproducibility research from NLG evaluations to data labelling for a supervised NLP classification task.

2.2. Understanding Sentiment in Code-Mixed Data

A major challenge for sentiment analysis is the ambiguities that hinder accurate classification. The classes and definitions can vary widely, which complicates the standardisation of this task across studies. Context and language can significantly affect the sentiment perceived. For example, Moore and Rayson (2018) show that identifying idioms, detecting sarcasm, and understanding the role of modifiers can influence the sentiment and the accuracy and replicability of the task. Typically, classifiers are designed only to process text that are written in high resource languages such as English. However, many other languages are used for digital communication (which is often the focus of a sentiment analysis task) and code-mixing is widely used in multilingual societies.

Malayalam (`ml`) is a Dravidian language[3] distinguished by its complex and rich phonetic and grammatical structures. Some research has been conducted on sentiment analysis for Malayalam text (Nair et al., 2014), for example focusing on tweets (Soumya and Pramod, 2020). Less research has been conducted on Malayalam-English code-mixed data, and a plausible reason for this is the lack of data availability. However, Chakravarthi et al. (2020) presented a corpus for sentiment analysis of code-mixed text for Malayalam-English, which we re-examine here.

The influence of code-mixing on annotator agreement and reproducibility has also received very little attention. One broadly related work is that of Abercrombie et al. (2023), who examined the impact of two factors, time and second language, on the inter- and intra-annotator agreement in German and English texts for a hate speech labelling task. Importantly, they found that label collection is not as repeatable as assumed even with the same annotators (in either language), which raises interesting questions on the reproducibility of multi-lingual data in general. In this study, we focus on what we believe is an understudied aspect of reproducibility in NLP, i.e. reproducibility of a sentiment analysis task using code-mixed data.

3. Chakravarthi et al. (2020): A Sentiment Analysis Dataset for Code-Mixed Malayalam-English

The original study by Chakravarthi et al. (2020) is comprised of data collection and labelling, as well as automated classification experiments. We provide a brief overview here.

3.1. Original Data Collection

In the original study, Chakravarthi et al. (2020) extracted 116,711 sentences from comments posted on YouTube about trailers for Malayalam movies from the year 2019, using the search term '*Malayalam movie 2019*', excluding instances that were in Malayalam script. The data gathered was then filtered to exclude any data that was non-code-mixed, i.e purely in English. The code-mixed content was then preprocessed, specifying that emojis were removed from sentences and sentences exceeding more than 15 words or fewer than 5 words were discarded. The resulting corpus contains 6,738 instances.

3.2. Original Annotation Process

The initial data labelling process was carried out by volunteer annotators. The label schema consisted of the following labels: *positive*, *negative*, *mixed feelings*, *neutral* and *non-Malayalam*. We follow the annotation process detailed in the original study which consists of three steps:

- **First Step:** Each item was labelled by two annotators independently. Items with the same two labels were considered finalised.
- **Second Step:** Items with label disagreements were annotated by a third annotator. Where agreement could be found among the three annotators, the labels were decided by majority vote (i.e. two out of three labels).
- **Third Step:** If there was no majority, these items were subsequently reviewed by two other annotators. Labels were again decided by majority vote.

As well as the three steps mentioned above, the original study omits to mention what was done with sentences still having label disagreements following the third step. These samples could have all five labels differing, or an absence of a majority label (i.e., two votes to two labels respectively and one vote to another). On enquiry, the authors responded that they discarded items on which there was no agreement after all three stages, and these are not included in the data that is made available in the original study.

[3] For an overview of the complex linguistic landscape of South Asia, including Dravidian languages, see Hock and Bashir (2016).

3.3. Original Classification Experiments

Chakravarthi et al. (2020) used a range of ML classifiers such as logistic regression (`LR`), support vector machines (`SVM`), Random Forests (`RF`), K-nearest neighbours (`KNN`), and multinominal naive Bayes (`MNB`) along with Term-Frequency Inverse Document Frequency (TF-IDF) for feature selection. Additionally, they also implemented four deep learning classifiers: 1D Dimensional Convolution (Zhou et al., 2016), Dynamic Meta-Embeddigs (`DME`), Contextualised `DME` (`CDME`) (Kiela et al., 2018) and `BERT` (Devlin et al., 2019). We select the two classifiers that attained highest performance (as detailed in Table 7), `LR` and `BERT`, for our classification reproduction study (see section 5).

4. Reproduction Study

4.1. Data

To maximise our resources, we selected only the test set split of the corpus for re-annotation.

4.2. Annotation Reproduction

We endeavoured to follow the original data annotation process as far as possible. However, to avoid discarding further items that were included in the original dataset, we incorporated a slight modification. For sentences that did not reach consensus among the initial five annotators (i.e., until the third step as described in subsection 3.2), rather than exclude such data, we added two further steps:

- **Fourth Step:** In scenarios where there was no clear majority, a sixth annotator was introduced to label the remaining items with label disagreements.

The reproduction study in theory should not elicit high disagreement amongst the annotators, as samples that did not have a majority label have already been discarded from the available data of the original study. Surprisingly, our re-annotation still yielded 61 items with label disagreements unresolved after step three, and 21 after step four. We therefore added a fifth step:

- **Fifth Step:** In case of unresolved disagreement among six annotators, such as an even split across three labels, ties between two labels, or a distributed disagreement (e.g., 2 *Positive*, 2 *Negative*, 1 *Neutral*, 1 *Mixed feelings*), the remaining 21 label disagreements were resolved by the first author of this work.

This process is detailed in Table 1.

Step	No. of Annotators	Labels collected	Disagreements
1	2	2	592
2	1	3	175
3	2	5	61
4	1	6	21
5	1	7	0

Table 1: Numbers of annotators, labels collected per item, and disagrements for each step.

4.3. Annotation Platform

We collected labels using the application `MS Excel`. The data was split into batches and was populated for annotation, which was undertaken on the annotators' personal computers.

4.4. Annotation Instructions

Chakravarthi et al. (2020) provide label definitions which they loosely adapt from Mohammad (2016), but little detail of the actual instructions given to the annotators. We contacted the first author for further clarification, but these were unavailable. We used the original study's label categories and definitions, and added an additional objective to instruct annotators on the task:

- **Objectives:** Categorise each sentence into one of the following segments.
  - **Positive:** There is an explicit or implicit clue in the text suggesting that the speaker is in a positive state, i.e., happy, admiring, relaxed, and forgiving.
  - **Negative:** There is an explicit or implicit clue in the text suggesting that the speaker is in a negative state, i.e., sad, angry, anxious, and violent.
  - **Neutral:** There is no explicit or implicit indicator of the speaker's emotional state: Examples are asking for like or subscription or questions about the release date or movie dialogue. This state can be considered as a neutral state.
  - **Mixed feelings:** There is an explicit or implicit clue in the text suggesting that the speaker is experiencing both positive and negative feeling: Comparing two movies
  - **Non-Malayalam:** For Malayalam if the sentence does not contain Malayalam then it is not Malayalam.

There were a total of 14 batches, where the first 13 batches had 100 items respectively and the last batch had 48 items to make a combined total of 1348 items.

Metric	Original	Updated
Language Pair	Malayalam-English	Malayalam-English
Number of Tokens	70,075	61,022
Vocabulary Size	19,992	19,389
Number of Samples	6,739	6,739
Number of Sentences	7,743	7,787
Average Sentence Length	10	8.26
Average Sentences Per Sample	1	1.15

Table 2: Comparison of corpus statistics reported by Chakravarthi et al. (2020) and our analysis.

Corpus	Before	After
Test Size	1,348	1,181
Train Size	4,851	4,283
Validation Size	540	463
Total Size	6,739	5,927

Table 3: Comparison of corpus partition sizes before and after preprocessing.

4.5. Preprocessing

While the preprocessing steps were outlined in the original study, it doesn't specify the packages used. Preprocessing is typically done before classifier training to prepare the data. However, according to the original study, the preprocessing phase was conducted before the data was made available to the annotators, as they illustrate in Figure 1, this was done to make annotation easier for the annotators. This motivation is unclear, as intuitively including emojis may provide more context, particularly for those examples where the sentiment is ambiguous. Data statistics are reported based on this preprocessed corpus. To confirm whether the provided data had already undergone the steps mentioned as per the original study we conducted the following preprocessing steps:

- **Removing emojis:** We removed emojis using the `emoji` package.
- **Sentence length adjustment:** We removed items with more than 15 words or less than 5 words with the `NLTK` tokeniser.

We maintain the test, train, and validation splits from the original study online.[4] However, after performing the preprocessing detailed above, the total number of samples have been reduced from 6,738 to 5,927. There were 309, 510, and 301 sentences that contained emojis, sentences exceeding 15 words and sentences fewer than 5 words, respectively. The data statistics are detailed in Table 3.

Figure 1 is taken from the original study, and shows that the preprocessing steps were conducted prior to the start of the annotation process. This means that the provided labelled data was expected to have undergone the process of removing emojis and sentences exceeding the sentence length criteria. The descrepancies observed in this post-preprocessed data indicate that there are deviations between the actual preprocessing and preprocessing steps reported in the original study, and this in turn raises consistency issues for the data we use in the reproduction study. Hence, we decided not to perform any preprocessing steps to preserve the same corpus size before commencing the comparative corpus analysis and feeding the data to the classifiers.

4.6. Comparative Corpus Analysis

Comparison of the original and updated corpus statistics are detailed in Table 2.

Analysis led to some observations that are slightly different from the original findings, possibly due to variations in tools used for preprocessing and analysis. These are outlined as follows:

- **Preprocessing:** Revisiting the earlier observations, the presence of emojis and sentences exceeding specified length criteria (before making the data available) highlights the preprocessing discrepancies that we found in the original study.
- **Corpus splits:** According to the original study, the corpus includes 6,739 comments or posts. This corpus was further divided into 20% for testing (i.e., 1,348), 10% for validation (i.e., 674) and remaining 70% for training. However, upon reviewing the data provided by the original study,[5] [6] we did not find this reported distribution. The data provided online has the following characteristics: while the test set contained the expected 20% of data (i.e., 1,348 items), the validation set had only 8.01% of data (i.e., 540 items), and the training set comprised 71.98% of data (i.e., 4851 items).

[4] https://dravidian-codemix.github.io/2020/datasets.html

[5] https://dravidian-codemix.github.io/2020/datasets.html

[6] https://github.com/bharathichezhiyan/MalayalamMixSentiment

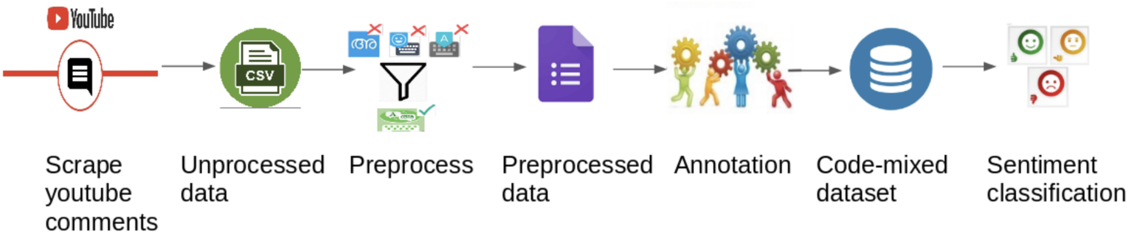

Figure 1: Data Collection Process Of Original Study (Chakravarthi et al., 2020)

- **Label imbalance:** A notable observation regarding this corpus is its imbalanced nature, with the distribution of labels heavily skewed. Specifically, the *Positive* and *Neutral* labels are significantly over represented with 41.71% of data (i.e., 2,811 items), and 28.24% of data (i.e., 1,903 items), respectively. The *Non-Malayalam*, *Negative* and *Mixed feelings* labels have only 13.12% of data (i.e., 884 items), 10.95% of data (i.e., 738 items), and 5.98% of data (i.e., 403 items), respectively. This imbalance could have implications for the performance of the sentiment analysis classifiers that are trained on this data, as they may be biased towards the more heavily represented labels.

- **Tokenisation:** We found 61,002 tokens in contrast to the 70,075 tokens reported in the original data statistics. This variation may be due to the differences in tokenisation process followed or due to the inclusion/exclusion of specific characters as tokens. We used the word and sentence tokenisers from NLTK.[7]

  We found 7,787 sentences, while the original data statistics reported 7,743. We found a vocabulary size of 19,389 compared to 19,992 reported in the original data statistics. The average sentence length observed was 8.26. These variations may be due to the differences in tokenisation processes or to the inclusion/exclusion of specific characters as tokens.

These observations do not diminish the value of the original corpus. Rather, they highlight the complexities and challenges of working with natural language, especially in a code-mixed environment.

5. Classification Models

To compare the effect of re-annotation on the downstream task, we reimplement two of the original supervised classification experiments.

The classifiers we apply are:

- **Logistic Regression (LR):** The choice to utilise LR was because of its simplicity and interpretability, and because it achieved some of the best results reported by Chakravarthi et al. (2020).

- **Bidirectional Encoder Representations from Transformers (BERT)** (Devlin et al., 2019): In addtition to its good performance in the original study, we used a multilingual BERT model due to its ability to handle the multilingual aspects of the corpus. The original study fails to specify the specific BERT classifier that was used. Given this lack of detail regarding the BERT classifier, we opted for `bert-base-multilingual-uncased`[8].

6. Results

We report results of the annotation reproduction study in subsection 6.1, and of automated sentiment classification in subsection 6.2.

6.1. Human Labelling

We observed notable shifts in the label counts across all labels, as shown in Table 4. There was an increase in the label counts of labels, *Positive*, *Negative* and *Mixed feelings*, and a decrease in those of the other two, *Neutral* and *Non-Malayalam*.

Label	Original	Re-annotated
Positive	565	626
Negative	138	162
Mixed feelings	70	144
Neutral	398	327
Non-Malayalam	177	89

Table 4: Comparison of original and re-annotated labels for each class for the test set.

The original corpus is reported to have a Krippendorff's alpha above 0.8, indicating a high level of agreement between the annotators across the whole corpus. However, our re-annoatation of the

[7] https://www.nltk.org

[8] https://huggingface.co/google-bert/bert-base-multilingual-uncased

test corpus yielded an alpha of only 0.383. This lower score signifies that there is notable annotator disagreement within the test corpus, highlighting the challenge of achieving label consistency. This disagreement can be seen in different rounds or steps of annotation as there were 592, 175, 61 and 21 label disagreements in the annotation process steps from one till four. Although these scores are not directly comparable due to the difference in size of the test corpus and the corpus as a whole, this outcome sheds light on potential inconsistencies in annotation reliability.

6.2. Classification Results

Original results We began by examining the performance of the LR and BERT classifiers reported in the original study. The outcomes of the original research are shown in Table 7, BERT achieving better preformance. The labels with the highest recall score for LR and BERT classifiers are, *Positive* and *Non-Malayalam*, respectively, suggesting its effectiveness in identifying those labels.

Reproduction results For both classifiers, we evaluated their performance on the original test corpus and our re-annotated test corpus, the evaluated performance of the classifiers are shown in Table 8.

When applying the re-implemented classifiers to the original test corpus, we observed similar results to that of the classifier reported by the original study, as seen in Table 5. This indicates that the re-implementation of both classifiers can be deemed successful, and the classifiers can now be utilised to conduct a comparative analysis on both corpora.

	Original		**Re-implemented**	
Classifier	LR	BERT	LR	BERT
macro	0.58	0.61	0.54	0.65
weighted	0.66	0.75	0.63	0.71

Table 5: Results obtained by re-implementing the 2 best classifiers using the original corpus, compared to the results given in Chakravarthi et al. (2020). Note, detailed results from the original work are given in Table 7.

The comparative analysis of the re-implemented classifiers on the original and re-annotated test corpora yielded the results that are detailed in Table 8. The analysis indicates that there is a decrease in the performance of both the classifiers, as seen in Table 6.

However, the BERT classifier suffered a greater decrease in performance. While LR relies on feature engineering and does not have any multilingual understanding capabilities, BERT is dependent on context and subtleties within the language, and might be more sensitive to modifications within the data like the labels that are assigned. This justifies the performance drop and implies that the BERT classifier has capabilities for capturing linguistic intricacies. This variance highlights the influence of annotation guidelines and newly annotated labels on the classifier performance.

	% Decrease	
Average	LR	BERT
macro	11.11%	20.00%
weighted	4.76%	14.08%

Table 6: Average F1-score decline: Classifier results on re-annotated vs. original corpus.

Furthermore, given the corpora's imbalance, with the *Positive* label having the highest number of instances, it can be observed this label has the highest recall rate among all labels for both classifiers and across both corpora, indicating effectiveness of the classifiers in identifying the sentences with a positive sentiment. On the other hand, the *Mixed feelings* label exhibits the lowest recall rate across both corpora, indicating that the classifiers struggle to identify sentences with mixed sentiment.

6.3. Quantified Reproducibility Assessment Results

We report the reproducibility results following Belz et al. (2022). We report Type I results via coefficient of variation (CV*) and Type III results via Krippendorff's alpha (α).

6.3.1. Type I Results

The comparison between the original classifier and the re-implemented classifier performance (on the original test corpus) was done using CV* (Belz, 2022; Belz et al., 2022). This was calculated using the F1-Scores of both classifiers as detailed in Table 9. Overall, the low CV* values for the macro and weighted averages of the F1-scores indicate **moderate reproducibility of the classifiers**.

Moving forward, the CV* of the re-implemented classifier performance on the original versus the re-annotated corpus was calculated, detailed in Table 10. Overall, the LR model demonstrated a CV* of 11.73 and 4.86, and the BERT model showed a CV* of 22.16, and 15.11, for the macro and weighted averages, respectively. In summary, these values suggest **less reproducibility regarding the data labels**.

6.3.2. Type III Results

To report Inter-Study Agreement assessment, the labels of the original test corpus and re-annotated

	LR				BERT			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
Mixed feelings	0.59	0.23	0.33	70	0.00	0.00	0.00	70
Negative	0.70	0.45	0.55	138	0.57	0.55	0.56	138
Neutral	0.65	0.65	0.65	398	0.73	0.79	0.76	398
Non-Malayalam	0.69	0.58	0.63	177	0.87	0.93	0.90	177
Positive	0.68	0.83	0.75	565	0.83	0.87	0.85	565
macro avg	0.66	0.55	0.58	1348	0.60	0.63	0.61	1348
weighted avg	0.67	0.67	0.66	1348	0.73	0.78	0.75	1348

Table 7: Results of the two best performing classifiers copied from Chakravarthi et al. (2020).

	LR (Original Corpus)				LR (Re-annotated Corpus)			
	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support
Mixed feelings	0.80	0.17	0.28	70	0.73	0.08	0.14	144
Negative	0.77	0.36	0.49	138	0.80	0.31	0.45	162
Neutral	0.66	0.60	0.63	398	0.58	0.64	0.61	327
Non-Malayalam	0.74	0.51	0.61	177	0.39	0.54	0.45	89
Positive	0.62	0.85	0.72	565	0.68	0.85	0.76	626
macro avg	0.72	0.50	0.54	1348	0.64	0.48	0.48	1348
weighted avg	0.67	0.65	0.63	1348	0.66	0.63	0.60	1348
	BERT (Original Corpus)				**BERT (Re-annotated Corpus)**			
Mixed feelings	0.42	0.44	0.43	70	0.42	0.22	0.29	144
Negative	0.68	0.51	0.59	138	0.69	0.44	0.54	162
Neutral	0.66	0.71	0.68	398	0.50	0.66	0.57	327
Non-Malayalam	0.81	0.75	0.78	177	0.36	0.65	0.46	89
Positive	0.77	0.79	0.78	565	0.78	0.72	0.75	626
macro avg	0.67	0.64	0.65	1348	0.55	0.54	0.52	1348
weighted avg	0.72	0.72	0.71	1348	0.63	0.61	0.61	1348

Table 8: Classifier performance on the re-annotated corpus compared to the original corpus. Note, results are reported on the test set given that our reproduction study focuses on the test set labels.

test corpus are compared by calculating Krippendorff's alpha (α). The results are $\alpha = 0.43$. This score indicates only a moderate agreement between the original and re-annotated labels, and further suggests that there is some variability in the label consistency of the data. A detailed discussion about the label consistency is given subsequently.

7. Discussion

Label Differences The comparison of the label distribution between the original and re-annotated corpora highlight the label differences, as seen in Table 4. The labels *Mixed feelings* and *Non-Malayalam* saw significant variation, with an addition of 74 items and a reduction of 88 items, respectively. The variation in the *Mixed feelings* label implies that the instructions of how to assess sentiment complexity in the guidelines is unclear. Similarly, the discrepancy in the *Non-Malayalam* label suggests that there is a possible confusion among the annotators as to what qualifies as code-mixed and purely content that is not Malayalam. For example, the following examples are instances with high disagreement among annotators:

Example 1:
Tamil and Telugu padam pole aayalo...
Don't kill malayalam movies reality
Its similar to Tamil and Telugu films...
Don't kill malayalam movies reality
Assigned Label: *Mixed feelings*

Example 2:
Numma or nummade or nammande palakkad le Katha aanu
Our own palakkads story
Assigned Label: *Neutral*

In **example 1**, there are three different perspectives. Firstly, the sentence could be seen as a ***neutral*** observation where Malayalam films are being compared to Tamil and Telugu films. Secondly, the advice '*Don't kill malayalam movies reality*' implies a ***negative*** sentiment towards the Tamil and Telugu industries. Thirdly, the sentence might imply a positive view towards Tamil and Telugu cinema's handling of reality and then warn against the destruction of reality in Malayalam films, suggesting a ***mixed sentiment***.

In **example 2**, the underlying sentiments are

	LR			BERT		
Labels	Original	Re-implemented	**CV***	Original	Re-implemented	**CV***
Positive	0.75	0.72	4.07	0.85	0.78	8.56
Negative	0.55	0.49	11.50	0.56	0.59	**5.20**
Mixed feelings	0.33	0.28	**16.34**	0.00	0.43	**199.40**
Neutral	0.65	0.63	**3.12**	0.76	0.68	11.08
Non-Malayalam	0.63	0.61	3.22	0.90	0.78	14.25
macro avg	0.58	0.54	7.12	0.61	0.65	6.33
weighted avg	0.66	0.63	4.64	0.75	0.71	5.46

Table 9: Quantitative Reproducibility Analysis utilising CV* between the results reported in the original paper and our re-implemented classifiers (using the original corpus). CV* is calculated based on the F1-scores.

	LR			BERT		
Labels	Original	Re-annotated	**CV***	Original	Re-annotated	**CV***
Positive	0.72	0.76	5.39	0.78	0.75	**3.90**
Negative	0.49	0.45	8.49	0.59	0.54	8.82
Mixed feelings	0.28	0.14	**66.47**	0.43	0.29	38.78
Neutral	0.63	0.61	**3.22**	0.68	0.57	17.55
Non-Malayalam	0.61	0.45	30.10	0.78	0.46	**51.46**
macro avg	0.54	0.48	11.73	0.65	0.52	22.16
weighted avg	0.63	0.60	4.86	0.71	0.61	15.11

Table 10: Quantitative Reproducibility Analysis utilising CV* of the re-implemented classifiers on the original versus the re-annotated corpus (i.e. detailed in Table 8). CV* is calculated based on the F1-scores.

Labels	**Example 1**	**Example 2**
Positive	0	2
Negative	2	0
Neutral	2	3
Mixed feelings	3	0
Non-Malayalam	0	2

Table 11: Comparison of labels assigned to example items.

positive, *neutral* and *non-Malayalam*. ***Positive*** because the phrase suggests pride to be part of the Palakkad district. Without context, the sentence could be seen as simply stating a fact, thus implying the ***neutral*** sentiment. Lastly, the code-mixed text can be interpreted as both Malayalam or Kannada, as '*Numma*' or '*Nummade*' are both words that are present in both languagues, this confusion can lead annotators to opt for the ***non-malayalam*** sentiment.

Moreover, the removal of emojis before annotation could have a significant effect on the underlying sentiment. Additionally, the challenges in code-mixed data such as the ambiguity outlined in the examples earlier could have been lessened with the help of more clear and detailed annotation guidelines.

Issues Affecting Reproducbility In the process of attempting to reproduce the results of another study, we faced several significant challenges that underscore the complexities of research reproduciblity. The following list outlines the reproduction challenges that were encountered:

- **Data Preparation Issues:** Chakravarthi et al. (2020) explain that *preprocessing* efforts were conducted to alleviate potential challenges for the annotators. However, the labelled data had numerous instances that appear not to have undergone preprocessing. The discrepancy between the documentation and the provided data poses a significant challenge to the reproducibility and hinders the integrity of the preprocessed data. Moreover, the study followed a structured approach to the *annotation process* which involved a three-step process. However, in this methodology, there is a critical ambiguity in addressing scenarios where the annotators continued in disagreement beyond the third step. Unlike the study's decision to discard such data, this reproducibility challenge was addressed by taking the decision to involve a sixth annotator to resolve those disagreements, and any other pending disagreements afterwards were resolved by me. Lastly, the absence of the actual *annotation guidelines*, apart from the basic schema, presented a significant challenge. Without these guidelines annotators faced ambiguity and had varied interpretations for the same sentences.

- **Classification Issues:** Although the total size of the provided corpus is accurate, the specified partition counts mentioned in the original study for training and validation is incorrect. This creates confusion and inconsistency in understanding the *corpus partitions*, which affects the reliability and reproducibility of the study and its corpus. Additionally, the original study asserts the free *availability of the code* and corpus for research purposes. However, this assertion is not met as the GitHub repository only contains a readme file with the corpus links but lacks the actual code. This situation complicates the replication process, stressing the importance of resource sharing in the NLP community. Moreover, acheiving comparable classifier performance given the lack of access to the original code, also posed a significant reproduciblity challenge. Furthermore, there is an uncertainity in the *classifier variant selection* for BERT in the original study. This oversight in not specifying the version was resolved by opting for the BERT-uncased-multilingual version. However, the differences in classifier version can hinder the results, thereby, affecting the reproducibility of the original study.

8. Conclusion

Our findings contribute to the ongoing discussion on the reproducibility and authenticity of research conducted in the field of NLP. The reproduction study yielded results that demonstrate a decrease in the performance accuracy of the re-implemented classifiers when compared to the results of the original study. Subsequently, **we were not able to reproduce the original study's results**. The obstacles we faced were **preprocessing inconsistencies**, **lack of guidelines and code**, **unclear annotation processes**, and **missing information** regarding packages and classifier variants used in the original study.

To advance the field and mitigate these reproducibility challenges, future work should focus on the development and adoption of reporting frameworks that are standardised. Additionally, the sharing of code, corpora, and detailed methodologies should be encouraged in the NLP community and studies assessing reproduciblity should be conducted systematically to pave the way for reliable and authentic researches.

Limitations and Ethical Considerations

Limitations Although this research provides insights into the reproducibility of NLP label collection, it has several limitations. The scope of this study is limited to the test corpus and set of preselected ML classifiers of the original research, which may not fully capture the underlying reproducbility challenges. Moreover, although the `bert-base-multilingual-uncased` classifier is designed to handle multiple languages, the study's approach, including the use of `LR`, neither the original nor this study explicitly addresses the intricacies of code-mixing. Furthermore, even though we were successful in re-implementing the classifiers in this study, it still might not mirror exactly those used in the original study, influencing the performance comparison and assessment of reproducbility.

Ethical Considerations This study was conducted with the approval of the institutional review board of Heriot-Watt University. Data was collected and stored on the Heriot-Watt-approved MS OneDrive system and complies with the General Data Protection Regulation (GDPR). Participant consent was obtained through an online information sheet and consent form prior to any data collection.

Acknowledgements

We thank the reviewers for their useful feedback. We also thank the annotation team for their dilligent work: Aashima Poudhar, Abin Paul, Mariya Sebastian, Muhammed Nafihe, Neethu Jose, and Tijo Thomas.

Gavin Abercrombie was supported by the EPSRC project 'Equally Safe Online' (EP/W025493/1), and Tanvi Dinkar was supported by the EPSRC projects 'AISEC: AI Secure and Explainable by Construction' (EP/T026952/1) and 'Gender Bias in Conversational AI' (EP/T023767/1).

Bibliographical References

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 96–103. Association for Computational Linguistics.

Aryaman Arora, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2022. Computational historical linguistics and language diversity in South Asia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1396–1409, Dublin, Ireland. Association for Computational Linguistics.

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454.

Anya Belz. 2022. A metrological perspective on reproducibility in NLP*. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Anya Belz, Maja Popovic, and Simon Mille. 2022. Quantified reproducibility assessment of NLP results. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16–28, Dublin, Ireland. Association for Computational Linguistics.

Anya Belz and Craig Thomson. 2024. The 2024 repronlp shared task on reproducibility of evaluations in nlp: Overview and results. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M. Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, Tanvi Dinkar, Ondřej Dušek, Steffen Eger, Qixiang Fang, Mingqi Gao, Albert Gatt, Dimitra Gkatzia, Javier González-Corbelle, Dirk Hovy, Manuela Hürlimann, Takumi Ito, John D. Kelleher, Filip Klubicka, Emiel Krahmer, Huiyuan Lai, Chris van der Lee, Yiru Li, Saad Mahamood, Margot Mieskes, Emiel van Miltenburg, Pablo Mosteiro, Malvina Nissim, Natalie Parde, Ondřej Plátek, Verena Rieser, Jie Ruan, Joel Tetreault, Antonio Toral, Xiaojun Wan, Leo Wanner, Lewis Watson, and Diyi Yang. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Penelope Gardner-Chloros. 2009. *Code-switching*. Cambridge university press.

Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

H.H. Hock and E. Bashir. 2016. *The Languages and Linguistics of South Asia: A Comprehensive Guide*. The World of Linguistics [WOL]. De Gruyter.

Douwe Kiela, Changhan Wang, and Kyunghyun Cho. 2018. Dynamic meta-embeddings for improved sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1477, Brussels, Belgium. Association for Computational Linguistics.

Gina-Anne Levow, Valerie Freeman, Alena Hrynkevich, Mari Ostendorf, Richard Wright, Julian Chan, Yi Luan, and Trang Tran. 2015. Recognition of stance strength and polarity in spontaneous speech. pages 236–241.

Bing Liu. 2012. *Sentiment analysis and opinion mining Bing Liu.* Synthesis digital library of engineering and computer science. Morgan & Claypool, San Rafael, Calif. (1537 Fourth Street, San Rafael, CA 94901 USA).

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

Andrew Moore and Paul Rayson. 2018. Bringing replication and reproduction together with generalisability in NLP: Three reproduction studies for target dependent sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1132–1144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mika V. Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32.

Deepu S. Nair, Jisha P. Jayan, Rajeev R R, and Elizabeth Sherly. 2014. Sentima - sentiment extraction for malayalam. *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1719–1723.

Nicolas P Rougier, Konrad Hinsen, Frédéric Alexandre, Thomas Arildsen, Lorena A Barba, Fabien CY Benureau, C Titus Brown, Pierre De Buyl, Ozan Caglayan, Andrew P Davison, et al. 2017. Sustainable computational science: the rescience initiative. *PeerJ Computer Science*, 3:e142.

S. Soumya and K.V. Pramod. 2020. Sentiment analysis of Malayalam tweets using machine learning techniques. *ICT Express*, 6(4):300–305.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common Flaws in Running Human Evaluation Experiments in NLP. *Computational Linguistics*, pages 1–11.

Janyce M. Wiebe. 1990. Identifying subjective characters in narrative. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Martijn Wieling, Josine Rawee, and Gertjan van Noord. 2018. Squib: Reproducibility in computational linguistics: Are we willing to share? *Computational Linguistics*, 44(4):641–649.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360, Edinburgh, UK. Association for Computational Linguistics.

Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. 2016. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3485–3495, Osaka, Japan. The COLING 2016 Organizing Committee.