# Experiment Design WS 2023/24
# Exercise 2 Group 1B

DRILEIDA HOXHA, TU Vienna, Austria
ARTUR OHANIAN, TU Vienna, Austria
VLADIMIR PANIN, TU Vienna, Austria
PHILIPP RETTIG, TU Vienna, Austria

This work is a reproducibility study of the paper "Should I visit this place? Inclusion and Exclusion Phrase Mining from Reviews" [1] by Gurjar et al. The experimental set up, initial conditions and problems of the reproduction are discussed. The differences are reported in a systematic fashion and improvements are proposed. Overall, the result could be reproduced fairly well, however some values showed some sizeable deviations. A major concern were missing means or deviances to properly confirm the findings.

**ACM Reference Format:**
Drileida Hoxha, Artur Ohanian, Vladimir Panin, and Philipp Rettig. 2024. Experiment Design WS 2023/24 Exercise 2 Group 1B. 1, 1 (January 2024), 7 pages. https://doi.org/10.5281/zenodo.10577845

## 1 INTRODUCTION

The paper [1] deals with the area of tourism data mining, addressing the limitations of existing automated itinerary planning systems by focusing on the nuanced constraints users face when selecting tourist spots. With the exponential growth in global tourism, the authors propose to mine inclusion and exclusion phrases from tourism reviews and classify them into 11 categories. These categories encompass factors such as age, claustrophobia, family-friendliness, crowd levels, food preferences, handicap accessibility, hygiene, parking, price, queues, and preferred visiting times. The authors contribute a dataset of 2303 phrases from around 2000 reviews, publicly available for further research. They explore the effectiveness of various deep learning models, including Conditional Random Fields (CRFs), Bidirectional Long Short-Term Memory networks (BiLSTMs), and BERT, in addressing the proposed tasks.

This report is part of the Exercise 2 of the course "Experiment Design for Data Science" offered by the Technical University of Vienna in the winter semester 2023/2024. The aim is to reproduce the results of the given paper in groups of four people. Special focus was posed on questions of reproducibility. Special consideration was put to whether enough information was provided, especially whether the code was provided, obtainment of statistically different results and whether the results obtained could stem from the same distribution as the presented ones.

Authors' addresses: Drileida Hoxha, TU Vienna, Austria; Artur Ohanian, TU Vienna, Austria; Vladimir Panin, TU Vienna, Austria; Philipp Rettig, TU Vienna, Austria.

Manuscript submitted to ACM

## 2 FRAMEWORK & WORKFLOW

Kindly, the authors of the paper [1] did provide their code in a public GitHub repository [1]. First of all, it has to be noted, that in the Task 2 in the *Level-2 Classification.ipynb* file the *glove.6B.200d.txt* was referenced altough it was not provided. This is a text file containing pre-trained word vectors in the GloVe (Global Vectors for Word Representation) format. The reason for not providing it was due to the large files size. Therefore it had to be added manually on the device, which however can't be changed due to the storage limitations of GitHub. For the reproducibility task, all the models mentioned in the paper [1] were run again, some could be run locally, while others had to be executed using Google Colab as the models required deviating amounts of computational resources.

## 3 REPRODUCIBILITY

The study explores two word embedding methods, GloVe and ELMo, and employs various sequence labeling and multi-class classification models, including CRFs, BiLSTMs, BiLSTM-CRFs, and BERT. Conditional Random Fields (CRFs) are used for tasks where contextual information influences predictions, while Bidirectional LSTMs (BiLSTMs) leverage bidirectionality to enhance accuracy in sequence modeling. BiLSTM-CRFs combine both BiLSTM and CRF networks to efficiently incorporate past input features and sentence-level tag information. BERT, a bidirectional Transformer-encoder model, achieves high accuracies in multiple NLP tasks and is applied in this study for sequence labeling and multi-class classification.

### 3.1 General Remarks

In the paper presented only a single value for each performance metric was provided. Neither mean values nor deviations have been reported. Therefore, statistical tests like a paired t-test could not be performed. Furthermore, in the code provided, the absence of random seeds has to be highlighted, which were added in the reproduction. These aspects could be improved in the provided paper.

### 3.2 Inclusion/exclusion phrase mining

*3.2.1 CRF + Glove.* Due to the error, related to the import of the keras module, which has the same name as the file, where the CRF model was defined, it was renamed in order to resolve this issue.

*3.2.2 BiLSTM + GloVe.* In order to run this file, adjustments on the code were necessary. A few modifications were made to the loadGloveModel function to address the UnicodeDecodeError. Therefore, the encoding='utf-8' parameter was added to the open function to explicitly specify the UTF-8 encoding.

To compile and execute the evaluation code (`evaluate.cpp`), the following command was used:

```
g++ evaluate.cpp -o evaluate_executable
```

- Added #include <cstdint> to include the necessary header file for uint (unsigned integer) type.
- Replaced uint with unsigned int for better compatibility and clarity.
- Initialized nExprPredicted and nExprTrue with 0 at the beginning of the testSequential function.
- Moved the declaration of loop variables (i and j) to the beginning of the corresponding loops.
- Added the calculation for the F1 score (result.f1) using precision and recall.

[1] https://github.com/omkar2810/Inclusion_Exclusion_Phrase_Mining

Possibly due to running the code locally, we were not able to produce the same results when rerunning again. This would generate some output like "0/r , 1/r" .

*3.2.3 BiLSTM + ELMo.* Due to the time-consuming nature of training the BiLSTM ELMO locally, it was transferred to and executed in the google colab environment. The original code was likely intended for older python versions than those available in the colab environment, leading to some compatibility issues with the original import statements. The issues were addressed by modifying the following commands:

- Changed 'from keras.models import Input' to 'from keras.layers import Input'
- Changed 'from keras.layers.merge import add' to 'from keras.layers import add'

Additionally, tf.random.set_seed() was utilized to set a seed, as consistent scores could not be generated without this.

*3.2.4 BERT.* To rerun the code provided on GitHub for the Bert models, several steps were necessary. First, due to the extensive training time, running the code in Colab was essential. Second, the code required minor adjustments:

- The installation command for the NVCC Jupyter plugin was updated
- A seed value was incorporated using 'torch.manual_seed' to ensure the ability to rerun the code.

### 3.3 11-class categorization accuracy results

*3.3.1 SVM.* The model was relocated to a distinct file to segregate it from the large LSTM models, facilitating faster reproduction, because in the original repo it is executed only after them, which is not optimal.

*3.3.2 XGBoost.* The model was also transferred to the same file as SVM to segregate it from computationally expensive models, thereby enhancing organization and efficiency in model management.

*3.3.3 BiLSTM + GloVe & BiLSTM-CNN + GloVe and BiLSTM Attn + GloVe.* For all three BILSTM models no serious challenges in reproducibility were encountered. Besides importing the file *glove.6B.200d.txt* only library updates were necessary. All import statements starting with'*keras*' were changed to '*tensorflow.keras*'. Furthermore, '*from keras.layers.convolutional import Conv1D, MaxPooling1D*' was modified to '*from tensorflow.keras.layers import Conv1D, MaxPooling1D*'. The results were fairly consistent. When using different seeds the models didn't display much deviation.

*3.3.4 BERT.* Due to the computational intensity, the BERT model was again executed with the help of Google Colab. Furthermore, the provided code was partially faulty. Therefore, corrections were applied in the initial dataset loading and formatting process, addressing inconsistencies between the filenames and column headings specified in the code and those actually present in the CSV file.

## 4 RESULTS

The reproduction of the experiment was mostly successful. Each model was run five times with different seeds and therefore for each value five results could be obtained. The mean of those values as well as their percentage deviations from the paper discussed are going to be provided in the appendix. Table 1 displays the mean of the reproduced values for the 'Inclusion/Exclusion Phrase Mining' task, while table 2 details the mean of the results for the '11-class Categorization Accuracy' task, as obtained from rerunning the code. In addition to the reproduced results, tables 3 and 4 provide a detailed comparison, showing the percentage differences between the mean of our reproduced results and the original findings reported in the paper.

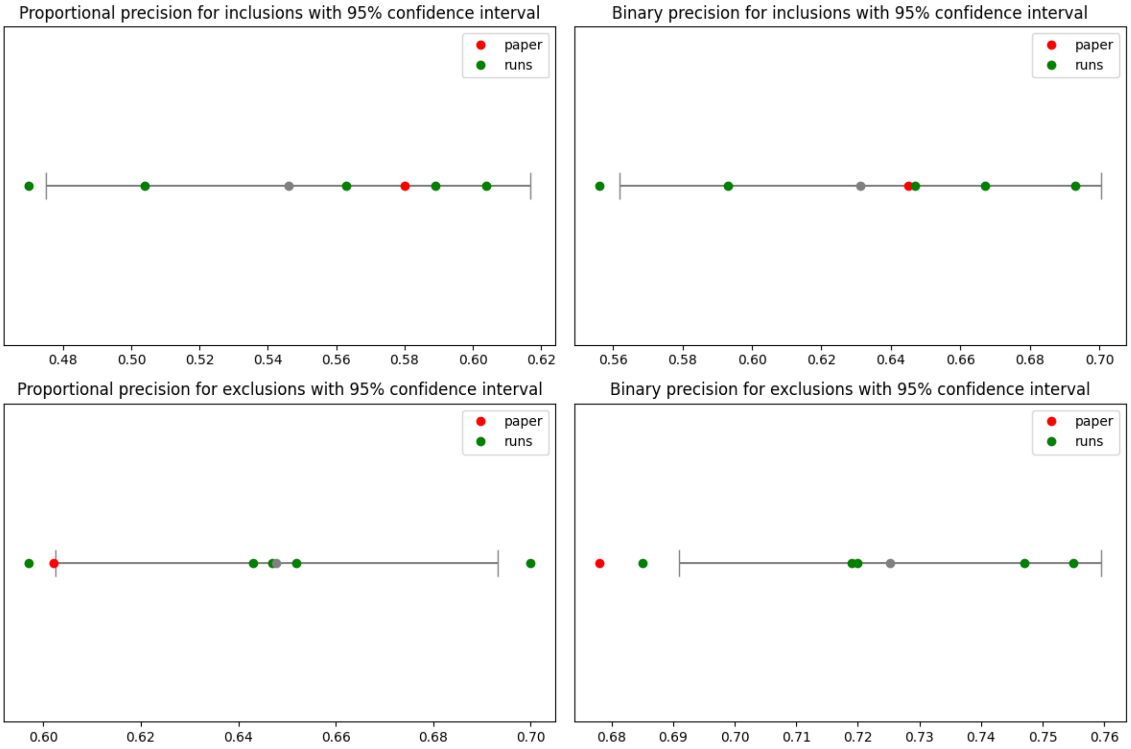

Fig. 1. Recall Scores for ELMO-LSTM Task 1

For the XGBoost model, the metrics were calculated in a different way, such that samples average values were taken instead of the accuracy.

Interestingly, in the results reproduced the accuracy score was the same for precision, recall and F1 score. However, in the paper slight deviations were observed, which is an interesting observation. In the 11-class categorization not all values from the classification report were given, like the weighted average or the macro average. In the paper no indication was given, why the other results were not revealed.

Difference in results can be explained by the different training splits and random seeds. Further, differences should not be relevant.

Figure 1 illustrates the process used for visually comparing the results from different runs with the metrics reported in the paper, in addition to the percentage differences. A 95% confidence interval was constructed based on the 5 different runs. This interval suggests that, under repeated sampling, the true mean of the performance metrics would be captured within such intervals about 95% of the time.

For most classifiers, the paper's reported values were either within the confidence interval or close to it. Notably, the models CRF + GloVe and BiLSTM + GloVe from task 1, and SVM and XGBoost from task 2, did not align well with the confidence interval. However, it has to be noted that a confidence interval derived from only 5 samples may not be a robust indicator of accuracy. Due to limited resources, it was not possible to opt for a more reliable sample size.

Figure 2 compares the average percentage differences, which shows minor deviations for the majority of the models, lying in a range of around 5%. Nonetheless, the graph suggests that it was not possible for us to reproduce the results exactly for the BiLSTM + GloVe classifer from task 1 as well as the SVM and XGBoost models from task 2.

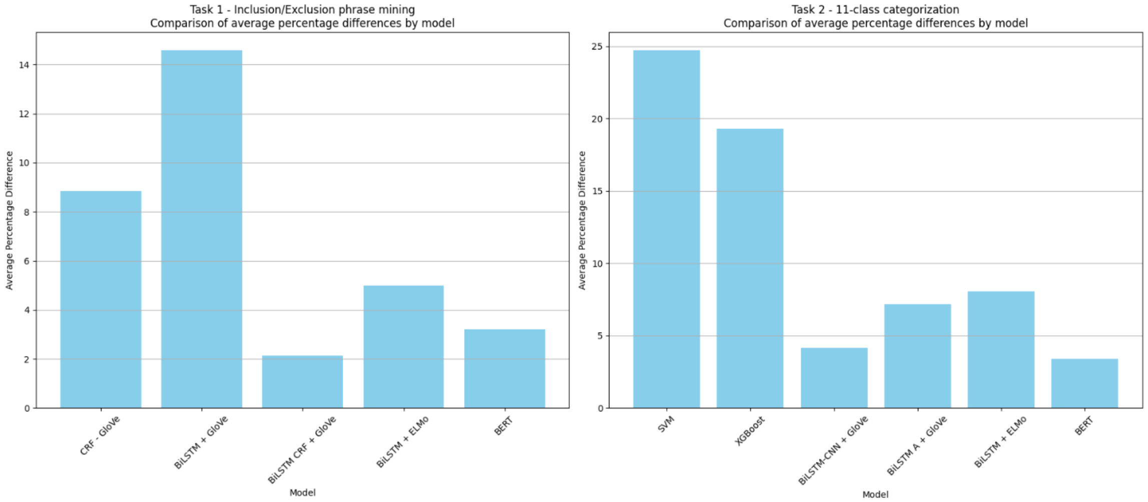

Fig. 2. Average Percentage Differences per Model

## 5 FINDINGS

In summary, this report successfully reproduced the majority of the results from the paper on inclusion and exclusion phrase mining from tourism reviews with minor deviations, roughly with an average of 5 percent. Despite this, some larger differences were observed for three models, with an average deviation between 15 and 25 percent. While some issues were observed, like library errors, due to the same naming for the repository file and keras model, word embedding files missing, or dependency issues, the key findings and trends in the paper's results were confirmed. We would suggest to the authors adding random seeds, providing confidence intervals and multiple values such that researchers can use statistical tests to ensure that the mean of the distributions obtained and reproduced are the same. Another recommendation would be to not name .py and .ipynb with the same name as the library modules in order to avoid namespace conflicts.

## REFERENCES

[1] Omkar Gurjar and Manish Gupta. 2020. Should I visit this place? Inclusion and Exclusion Phrase Mining from Reviews. arXiv:2012.10226 [cs.IR]

## A APPENDIX

| Model | Inclusion | | | | | | Exclusion | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | | Recall | | F1 | | Precision | | Recall | | F1 | |
| | Prop | Bin | Prop | Bin | Prop | Bin | Prop | Bin | Prop | Bin | Prop | Bin |
| CRF + GloVe | 0.34 | 0.38 | 0.55 | 0.74 | 0.42 | 0.50 | 0.29 | 0.32 | 0.52 | 0.71 | 0.37 | 0.44 |
| BiLSTM + GloVe | 0.51 | 0.54 | 0.23 | 0.54 | 0.31 | 0.54 | 0.67 | 0.69 | 0.09 | 0.23 | 0.16 | 0.33 |
| BiLSTM CRF + GloVe | 0.47 | 0.61 | 0.61 | 0.69 | 0.53 | 0.65 | 0.32 | 0.4 | 0.66 | 0.77 | 0.43 | 0.53 |
| BiLSTM + ELMo | 0.61 | 0.65 | 0.53 | 0.74 | 0.57 | 0.69 | 0.60 | 0.65 | 0.50 | 0.68 | 0.55 | 0.67 |
| BERT | 0.67 | 0.73 | 0.76 | 0.89 | 0.71 | 0.80 | 0.63 | 0.72 | 0.75 | 0.86 | 0.69 | 0.78 |

Table 1. Mean of Inclusion/exclusion phrase mining results

| Model | Total | | | Inclusion | | | Exclusion | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| SVM | 0.81 | 0.82 | 0.82 | 0.77 | 0.77 | 0.77 | 0.86 | 0.86 | 0.86 |
| XGBoost | 0.64 | 0.64 | 0.64 | 0.62 | 0.64 | 0.63 | 0.66 | 0.67 | 0.66 |
| BiLSTM-CNN + GloVe | 0.86 | 0.86 | 0.86 | 0.88 | 0.88 | 0.88 | 0.83 | 0.83 | 0.83 |
| BiLSTM A+ GloVe | 0.85 | 0.85 | 0.85 | 0.87 | 0.87 | 0.87 | 0.83 | 0.83 | 0.83 |
| BiLSTM + ELMo | 0.82 | 0.82 | 0.82 | 0.77 | 0.77 | 0.77 | 0.86 | 0.86 | 0.86 |
| BERT | 0.95 | 0.95 | 0.95 | 0.98 | 0.98 | 0.98 | 0.91 | 0.91 | 0.91 |

Table 2. Mean of 11-class categorization results

| Model | Inclusion (% Difference) | | | | | | Exclusion (% Difference) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | | Recall | | F1 | | Precision | | Recall | | F1 | |
| | Prop | Bin | Prop | Bin | Prop | Bin | Prop | Bin | Prop | Bin | Prop | Bin |
| CRF + GloVe | -5.08 | -9.11 | 3.01 | -2.51 | -2.12 | -6.88 | -23.92 | -19.90 | -0.38 | -2.61 | -15.63 | -15.04 |
| BiLSTM + GloVe | 3.95 | 3.05 | 5.76 | 7.31 | 4.72 | 5.20 | -37.15 | -36.99 | 16.67 | 15.87 | -19.78 | -18.62 |
| BiLSTM CRF + GloVe | -1.7 | -8.1 | -0.7 | -2.4 | -1.4 | -2 | -1.98 | -2.48 | -1 | -0.4 | -1.47 | -1.84 |
| BiLSTM + ELMo | 5.00 | 1.09 | -11.75 | -3.90 | -3.73 | -1.14 | -0.17 | -3.69 | -11.48 | -7.32 | -5.53 | -4.98 |
| BERT | -1.03 | -2.27 | -0.13 | 2.30 | -0.56 | -0.12 | -5.12 | -4.76 | -6.37 | -5.29 | -4.96 | -5.45 |

Table 3. Percentage Differences in Inclusion/Exclusion phrase mining results

| Model | Total (% Difference) | | | Inclusion (% Difference) | | | Exclusion (% Difference) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| SVM | 12.14 | 29.32 | 30.67 | 1.71 | 21.57 | 18.95 | 29.02 | 37.06 | 42.05 |
| XGBoost | -20.70 | -19.35 | -18.95 | -22.19 | -18.60 | -19.97 | -19.46 | -16.75 | -17.87 |
| BiLSTM-CNN + GloVe | -4.47 | -4.15 | -4.04 | -2.10 | -1.78 | -1.78 | -6.75 | -6.12 | -6.12 |
| BiLSTM A+ GloVe | -7.44 | -7.14 | -7.14 | -7.25 | -6.85 | -6.85 | -7.72 | -6.99 | -7.20 |
| BiLSTM + ELMo | -8.09 | -7.57 | -7.47 | -16.18 | -15.81 | -15.72 | -0.46 | 0.70 | 0.59 |
| BERT | -3.27 | -3.27 | -3.27 | -0.81 | -0.71 | -0.71 | -6.26 | -6.06 | -6.06 |

Table 4. Percentage Differences in 11-class categorization results