SHOW WE DO A TOPIC BASED GLOSSARY¿?¿?

1

2

- recycle some of the state of the art discussion, colab bt pyscho and ml, techniques

3 Abstract

4 Introduction

5 Objectives

6 Data

7 Methodology

ALBERTO

[tie in with data section when talking about the reasons of the methodology chosen]

[do actual notation for the models?]

[https://arxiv.org/pdf/1810.04805 bert paper in case i cite]

[400 words give or take]

The literature review previously conducted proved the novelty of the intersection between the Rorschach test and natural language processing methods. At the wake of this situation, we thought it beneficial to approach the task in a hierarchical way, attempting simpler and more explainable solutions first before moving on to more complex, back-box strategies.

Given the scarce nature of Rorschach related data and the heavy class imbalances present in it, it seemed plausible at the start of the research that training models with the original labels would not produce a satisfying result. In an attempt to remedy this would-be set back, we developed two sets of models parallelly: one set whose determinant and content labels are the canonical ones provided originally and a second set of models whose labels would be reduced by grouping. The grouping was carried out by semantic proximity while simultaneously aiming to balance the frequency of the classes. [we will have to talk about his tradeoff in the data section]

Once these two data approaches were defined, the next step was deciding which machine learning techniques to apply. After the initial failure of the traditional algorithms implemented [link to the mitigation with a section talking about it] and the impossibility to acquire enough data to train a neural network from scratch, it was clear that leveraging transfer learning approaches was the way ahead [dont like this, look for synonym]. Fine-tuning pre-trained, task-agnostic BERT models, more specifically, the bert-base-uncase [cite] model SUBJECT TO CHANGE!!!!!!, remedied the shortage of domain-specific data while achieving much better results than the first batch of models.

After deciding on this model architecture, gridsearch was conducted to find the optimal hyperparameters for the final models. Because of the computing-devouring nature of such process, the search was organized in three stages to minimize its environmental impact: an initial exploration first, a second search where the number of hyperparameter options was significantly cut down and a final stage where the best two to three models for each data representation were trained longer. With the gridsearch concluded, a model was picked for each data approach for each label type, that is, one for the individual determinant labels, one for the individual content labels, one for the grouped determinant labels and one for the grouped content labels, and the models were compared for a final evaluation. [WE HAVENT DONE THIS YEt]

8 Evaluation

9 Future Work

ALBERTO [open source data sheet with info about psychology related datasets can be found in https://osf.io/th8ew/ Use this to justify the lack of data?] some random ideas for future work:

[https://openai.com/index/hello-gpt-4o/]

[360]

A common theme discussed throughout this research is that of data quantity and quality. In order to produce more robust and reliable models, regardless of the architecture chosen to develop them, a corpus building project needs to be carried out not only to collect anonymized Rorschach protocols, but also to develop a consistent annotation guideline to deal with the ambiguity and subjectivity of the test. Given the technical nature of the coding process it is crucial that trained psychologists experienced in the Rorschach are at the center of this corpus building process.

Regarding a potential continuation of the line of work presented in this research, a logical next step could be to merge the two approaches. If each approach is taken to be a function that classifies a certain input to a macro-category, in the case of the grouped labels, or a proper class canonical to the test, in the other case, a composition of both approaches could be conceived as a function that assigns the input to however many macro-categories first and, depending on the macro-categories output, assign one canonical label for each macro-category detected. Such an approach would leverage the higher reliability of the grouped-labels approach without the need to sacrifice the specificity of the second approach.

While the Rorschach coding was strategysed as a multi-label classification problem in this research, other strategies that fullfull the objectives of the underlying problem to tackle, coding the test swiftly and reliably, should be acknowledged as well. Just a few days before the writing of this paper, OpenAI presented its latest flagship model, GPT-4o, leaving the general public in awe over the smooth human-computer interaction experience it provided in its demo. Adding the rising trend of retrieval augmented generation [add source] to the equation, it is no longer hard to imagine an assistant-like agent that collaborates with a domain expert [add source liu et al] and that provides suggestions based on ground truth minimizing possible hallucinations.

https://aclanthology.org/2023.ccl-2.2.pdf
https://research.ibm.com/blog/retrieval-augmented-generation-RAG
https://arxiv.org/pdf/2005.11401v4

All in all, whichever direction the relationship between the Rorschach and NLP may take, the continued collaboration between psychology and NLP practitioners will be crucial to the development of satisfying solutions. [add more stuff about encouraging psy ppl to attend technical/human in the loop foundational formation so that tehy understand the processes???]

10

11 Machine Learning Models

As discussed in the Methodology [href to the method section], simpler models were developed prior to the transformer models. The small size of the data allowed for the exploration of a wide variety of algorithms without having to worry about computing consumption. These first experiments consisted of the following machine learning algorithms: logistic regression, support vector machine, stochastic gradient descent, k-nearest neighbors, random forest and gradient boosting classifier.

The initial models followed the two data approaches previously defined. As far as the text vectorization techniques implemented, one set of models used tf-idf [add explanation of tf somewhere] after pre-processing the text removing stopwords and another set of models used the centroid of the array resulting from the tokenization of the text using the all-MiniLM-L6-v2 [add citation] language model.

- Let $label_{determinants}$,$label_{contents} \in L$ be the set of labels type that has to be classified.
- Let d_{macro},$d_{individual} \in D$ be the set of possible data approaches.
- Let a_{logreg},a_{svc},a_{sgd},a_{knn},a_{rf},$a_{gdc} \in A$ be the set of algorithms used for the model development.
- Let $vect_{tf-idf}$,$vect_{token} \in V$ be the set of text vectorization techniques applied.

Then the set of models developed is defined as $M = L \times D \times A \times V$.

11.1 Results

In general, the results for the models were underwhelming with a few exceptions. The analysis of the results will be split by the vectorization technique used for each set of models, the main focus will be observing the trends by data approach rather than centering on individual models and their potential use and the f1 score will be the main metric used for the evaluation.

11.1.1 Tf-idf Models

[add explanation]

Table 1: Mean F1 score per data approach

Data approach	F1 score
Individual labels	0.15
Macro labels	0.41

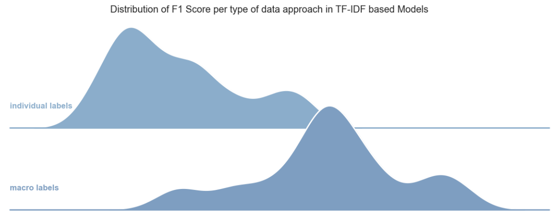
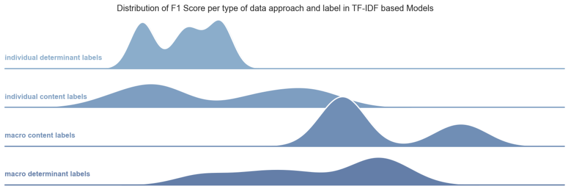


Table 2: Mean F1 score per data approach and label

Data approach and label	F1 score
Individual content labels	0.18
Individual determinant labels	0.11
Macro content labels	0.46
Macro determinant labels	0.35



11.1.2 all-MiniLM-L6-v2 Models

[add explanation]

11.1.3 Conclusion

[stat test for macro vs individual???]

Table 3: Mean F1 score per data approach

Data approach	F1 score
individual labels	0.08
macro labels	0.21

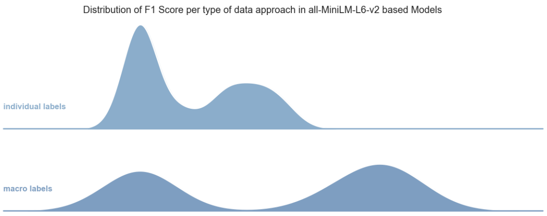
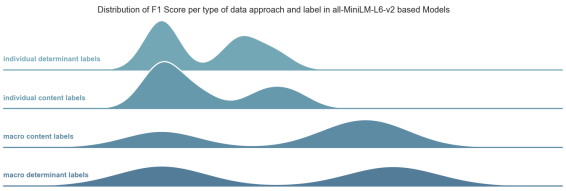


Table 4: Mean F1 score per data approach and label

Data approach and label	F1 score
Individual content labels	0.08
Individual determinant labels	0.08
Macro content labels	0.23
Macro determinant labels	0.20


Distribution of F1 Score per type of data approach and label in all-MiniLM-L6-v2 based Models
individual determinant labels
individual content labels
macro content labels
macro determinant labels