Project Athena

Davide Borghini
d.borghini3@studenti.unipi.it

Davide Marchi
d.marchi5@studenti.unipi.it

Giordano Scerra
g.scerra1@studenti.unipi.it

Andrea Marino
a.marino47@studenti.unipi.it

Yuri Ermes Negri
y.negri@studenti.unipi.it

June 3, 2024

Abstract

Final report of Project Athena, the exam project of the HLT course, a.y. 2023-2024. Project Athena consists in text classification over sentences taken from philosophical books.

In the report the procedures that were followed to explore the dataset, as well as a contrastive analysis of such corpus against three other corpora, are outlined. The choice of the reference metric that was used to compare models is motivated. While working on the project, it was noted that the dataset may have inherent problems hampering the task of classification. The hypothesis that was formulated in that regard is described, as well as its assessment strategy.

The chosen models for comparisons are listed and described, their performance is reported. The achieved results are commented.

Contents

1 Introduction

Ah, philosophy! All about "why are we here?" and "what is our purpose?". Our purpose is to infer the school of thought behind every sentence we come across. We aim to do this by training on a multinomial text classification task different models over a dataset composed of philosophical writings taken from Project Gutenberg. In this report, we'll explain how the dataset is composed, what kind of models we worked with and why, what metric we chose to measure their performances, our hypothesis on how different training strategies would work, our experiments and their results.

2 Related works

The Philosophy Data Project is both a Dataset and a Data Analysis inquiry made by Kourosh Alizadeh, a Philosophy teacher and Data Analyst. Works related to the task of analyzing philosophical thought with this particular dataset were made by different authors (including Alizadeh himself) and published in the form of online notebooks on Kaggle [1].

3 Dataset

We conducted an in-depth exploration of the dataset, as well as a contrastive analysis. Both analysis focused on presence and distribution of words.

3.1 Dataset exploration

In the version of the dataset that we used [1] there are 360808 sentences in total. The sentences are taken from 59 texts from 36 authors, spanning 13 schools of thought (aka classes). The classes are extremely unbalanced, as the Figure 1 shows.

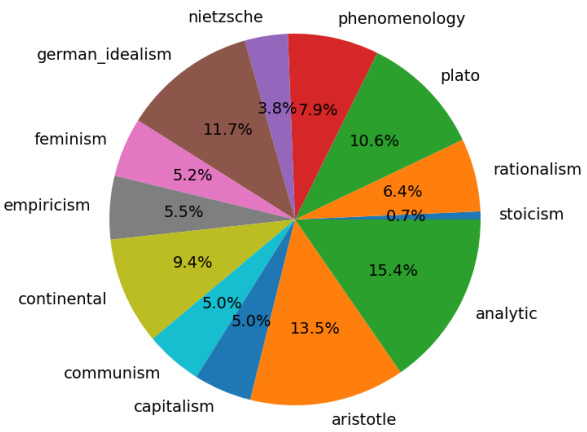
Basic statistics

We started with considering some basic statistics. In particular, the overall sentence length distribution was computed. We noted that it's very non-uniform – as the Figure 2 shows – and it's skewed towards short sentences (almost 30% of the sentences are less than 15 words long).

Words distribution

After these basic statistics, we considered the distribution of words inside our dataset. We created bag of words for each class, removing stopwords and non-alphanumeric characters, and putting all the words to lowercase. For this task, we considered the `tokenized_txt` column of the dataframe. The

Figure 1: *Class balancing in the Philosophy dataset*



content of this column has been obtained by the original curator of the dataset by applying Gensim's `simple_preprocess` to the sentences in `sentence_str`. In turn, the sentences in this column have been obtained through both a custom ad-hoc cleaning of the texts and spaCy `en_core_web_lg` model.

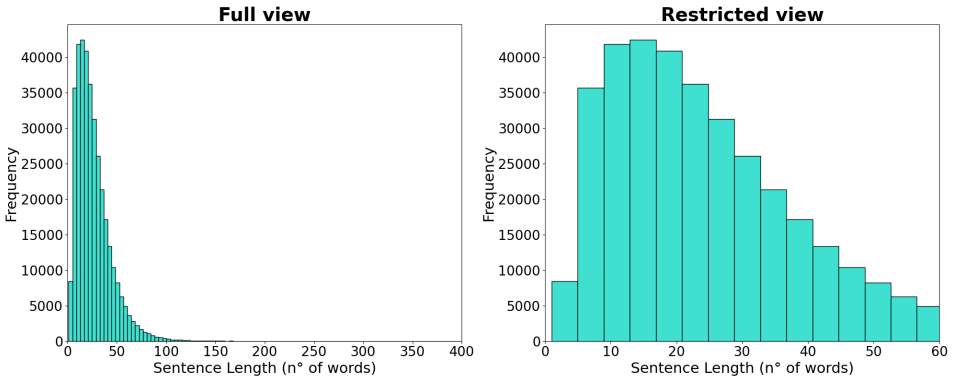
The stopwords make up from 48.88% (Feminism class) up to 55.79% (Plato class) of the total amount of words.

From the bag of words, we created wordclouds, heatmaps and histograms. They're available on the GitHub repository [3], with a short discussion.

3.2 Contrastive analysis

We compared our Philosophy corpus with these three corpora: Gutenberg [6], Brown [5], Simple English Wikipedia [8].

Figure 2: *Distribution of sentence lengths in the Philosophy dataset*



The contrastive analysis we propose consists of three pairwise comparisons, each between the Philosophy corpus and one of the three aforementioned corpora.

Creation of the compared distributions

After a preliminary preprocessing and tokenization phase, the word distributions to be compared had to be created. This has been done through the following steps:

1. Stopwords and non-alphabetic character removal. For this, NLTK's list of stopwords was used

2. Thresholding. For each corpus, we considered only the top 80% of the words, ordered by frequency (75% for Simple English Wikipedia).
   In all of the cases, the top 80% is populated by very few words, compared to the overall size of the vocabularies (always between 3 and 6 thoudands)

3. Normalization. We normalize the count of the words to get values that would sum to 1.

Distribution comparison

The described steps return a distribution for each corpus that serves as a basis for the comparisons. The comparisons have been conducted in a similar way.

The first step is always the individuation of the common vocabulary (aka "shared support"). The two considered corpora will share part of the vocabulary, but the words in common may have very different frequency. The rest of the analysis is based on this crucial phase, and is made of the following steps:

- The sum of the frequency of the shared words is computed for each corpus' distribution, separately. This is a measure of "how much they have in common"

- The euclidean distance and the cosine similarity between the distributions' restriction on the shared support are computed. These are measures of "similarity".

Some results are summarized in the table 1.

The comparison with the Brown corpus shows that the Philosophy corpus' lexicon is peculiar. In fact, we got varying scores for the shared mass percentage and cosine similarity across the various categories. Our metrics allowed us to detect differences across the distributions. That is, our corpus is more comparable to some genres and less to others with respect to these

Table 1: *Results of the contrastive analysis*

Corpus	**Shared mass (%)**		**Euclidean distance**	**Cosine similarity**
	Ref. corpus	Philosophy		
Gutenberg	76.42%	73.17%	3.556×10^{-2}	0.5900
Brown (Belles Lettres)	76.60%	86.60%	1.905×10^{-2}	0.8445
Brown (Humor)	64.97%	57.24%	2.369×10^{-2}	0.7472
Simple English Wikipedia	62.91%	78.90%	3.001×10^{-2}	0.5664

measures. The most similar categories (Belles lettres and Lore), though, are related to philosophy in a very loose way, and it's not surprising that philosophy has little to do with humour. In all cases, the shared words are also the most frequent ones. More insight is available at [3].

Moreover, the results we got for the metrics are still pretty concentrated (cosine similarity for Gutenberg and Simple English Wikipedia, for example). Therefore, more aspects besides the lexical one must be considered, and more sophisticated models are necessary.

4 Chosen metric

To compare and evaluate the various models, we chose F_1 score macroaverage. This is mainly for two reasons. Firstly, our dataset is not well balanced, and the distribution of classes that we have is not to be considered representative of the real word; secondly given our task we think that false positive and false negative should be equally weighted.

5 Our hypothesis and its assessment strategy

Many short sentences in the dataset might come from different schools and don't even look well segmented. The longer sentences present in the dataset, on the other hand, could convey a whole lot of meaning on their own. What we think is that, generally, shorter sentences and the meaning they convey could be present in more than one philosophical work.

This has two consequences. Firstly, short sentences act in a way that is akin to stopwords, making them superfluous or outright disruptive with respect to the process of learning how to classify different philosophical currents and ways of thinking. Secondly, they degrade the score of any classifier, since they're inherently hard to be classified, whatever the label, because they convey less informational content.

Therefore, supposing our hypothesis is true, short sentences either act as noise or are irrelevant for our learning process. Considering the extensive analysis of the dataset we made in the previous sections, as well as the intermediate results, we decided to assess this hypothesis by training and validating all models on both the whole developement set and a reduced version of the developement set.

The reduced version of the developement set is obtained by trimming off a portion of the sentences that span from 0 to 14 words from the already splitted train/validation data. To do so, we computed the mean character length of these sentences, which we found to be $\approx 82, 71$. Hence we rounded to 83 and excluded all sentences whose length *in characters* is strictly less than 84.

What we get from this further splitting of the dataset is the set of relations defined below:

	≥ 84	Full
≥ 84	α	β
Full	γ	δ

where the rows define the models trained on either the full or the reduced dataset and the columns the validation over the full or reduced dataset. Considering what we said, what we expect is:

$$\alpha > \beta; \quad \beta > \delta; \quad \alpha > \gamma; \quad \gamma > \delta; \quad \alpha > \delta$$

6 Experiments (models)

The models we considered cover the three main classes of NLP models: generative (Naive Bayes), discriminative (Recurrent Neural Networks) and Transformer-based (BERT, DistilBert, Zero-shot `bart-large-mnli`).

6.1 Naive Bayes

This is the simplest model that we decided to try. At the beginning we thought this model would have been just a baseline, something fairly simple to beat, though we obtained incredibly good results with little effort and computational power. (For this same reasons this model is the one the author of the original dataset considers the best one).

We tried many variation of the basic settings: a Naive Bayes based on TF-IdF weighted term-document matrix, and another model in which negation token were prepended to words that were in a negative context. But in the end, the naivest of the models was among the best: The most important thing that improved the results was the tuning of the Laplace smoothing hyperparameter.

6.2 Recurrent Neural Networks

We followed different approaches that approximately gave the same results: we both learned the embeddings from scratch during training and used pretrained GloVe vectors; we used Long Short Term Memory units, Gated Recurrent Units and normal Dense layers. As the hidden layer, we tried LSTMs, GRUs and simple RNN cells. What worked best for us was using a Bidirectional LSTM layer with 64 units.

6.3 BERT

BERT [4] is the biggest model we trained. All of our phrases except one (that we decided to discard) were short enough to be passed to the model without the need for truncation.

The implementation of this model was straightforward thanks to the `transformer` library, which provide functions to obtain the BERT model with some layer already on top of the `[CLS]` token. We therefore have a classifier we can fine-tune out of the box.

6.4 DistilBERT

DistilBERT is a lighter and faster version of BERT [9]; so all there is to say about this model was already said in the above section. We decided to try out also this model to understand if it was really essential to use a huge and expensive model like BERT.

6.5 Zero-shot Learning with `bart-large-mnli`

Zero-shot classification categorizes texts into predefined classes, without requiring fine tuning with labeled training data for those specific classes. We used the `bart-large-mnli` [2] by Facebook for this task, chosen for its generalization capabilities from pre-trained knowledge.

Even implementing descriptive labels, the results were definitely underwhelming, and the ones we had the best results were formatted like: *Analytic Philosophy*, *Aristotelian Philosophy*, *German Idealism Philosophy*...

The vast amount of classes and the subtle similarities between them proved to be a too tough task for a not fine tuned approach.

7 Results

The validation scores of the models trained on the full dataset are consistently better than those of the models that were trained on the reduced dataset. From this, we conclude that training on the full dataset was the better choice, and that the shorter sentences are actually helpful in creating a better model. So, the shorter sentences can't be excluded without degrading performance

(on both versions of the validation set). In particular, for what concerns β and γ, this indicates that the "knowledge" acquired on the longer sentences is not a satisfying proxy for the one acquired on shorter sentences.

The short sentences are the hardest to classify, in fact both versions of each model achieve a better score on the reduced version of the validation set.

BERT and DistilBert perform similarly, and both significantly outperform all the other models. So, transformers are the best models for this task but – as we assumed – a large transformer is not necessary.

Table 2: F_1 scores of the compared models

Model	`BERT`		`NB`		`RNN`		`Zero-shot`		`DistilBert`	
TR \ **VL**	≥ 84	Full	≥ 84	Full	≥ 84	Full	≥ 84	Full	≥ 84	Full
≥ 84	0.86	0.80	0.80	0.75	0.77	0.72	0.12	0.13	0.86	0.80
Full	0.89	0.84	0.81	0.76	0.79	0.74	0.12	0.13	0.87	0.82

Due to constraints in terms of time, hardware resources, and environmental resources available on our planet, the best model – BERT trained on the full training set – has not been re-trained on the developement set before evaluating it on a previously held-out test set. Following the same schema described previously, it achieved an F_1 score of 0.83 on the full test set, and an F_1 score of 0.88 if the short sentences are excluded.

8 Conclusions

We showed the main steps of the exploration of the Philosophy dataset and of the performed contrastive analysis. We presented some of the conclusions therein. We described the chosen metric for model evaluation and comparison, motivating our choice.

We hypothesized that our dataset, besides the extreme class unbalancing, may have a problem would hinder the creation of a satisfactory classifier, that is the presence of short and unmeaningful sentences. We described how we assessed this hypothesis, and explained which results we expected if our hypothesis was true.

Ultimately, the results weren't in line with our predictions, but some results are in line with the phenomenon that we described. Moreover, the fine-tuned transformer based models did well on the task, but not staggeringly more than the simple Naive Bayes.

Therefore, this project could be further developed with more advanced data analysis techniques, which would allow us to better understand the composition of our dataset. It could be further developed with different data

selection strategy (over/undersampling), with the training of more advanced and bigger models or – better – a more careful training of simpler models.

A thing that we want to highlight is that we will not always be able to use the best model, because in our case the best ones are also the most expensive ones to run. If we think for example about the task of assigning a philosophical current to a book, we may want to analyze many of its phrases, and, since NB already reaches good performances, using an heavy model like BERT may not be worth.

So, ultimately we improved the results of Kourosh Alizadeh by a good margin, but at what cost? We estimated that, for the transformer-based models, we used the remote machine for 29 hours. This means that we emitted 3,76 kg of CO_2eq.[1]

The last thing we think is important to underline is that the classifiers we built can't actually tell to which philosophical current a sentence belongs to. This doesn't mean that our models don't work properly, this means that we're "only" able to classify the topic of a phrase. We can say that a phrase revolves around a concept that is discussed by a particular current, but if we give that same sentence negated to our model it will likely yield the same class since the opposite class is not among the other 12 of our dataset.

[1]We considered only the usage of the GPU needed to train and validate the three transformer-based models. Estimates were done using [7]. This number should only be considered a loose approximation since we've only taken into account part of our work and we don't know the carbon efficiency of the infrastructure we used.

References

[1] Kourosh Alizadeh. *History of Philosophy: Sentences taken from 51 texts spanning the history of philosophy*. Accessed: March 2024. kaggle. 2021. URL: https://www.kaggle.com/datasets/kouroshalizadeh/history-of-philosophy.

[2] *bart-large-mnli by Facebook*. Accessed: May 2024. HuggingFace. 2023. URL: https://huggingface.co/facebook/bart-large-mnli.

[3] D. Borghini et al. *Project Athena*. GitHub. 2024. URL: https://github.com/giordanoscerra/ProjectAthena.

[4] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[5] W. N. Francis and H. Kucera. *Brown Corpus*. Available at: https://www.nltk.org/nltk_data/. Brown University. July 1979. URL: http://korpus.uib.no/icame/manuals/BROWN/.

[6] *Gutenberg Corpus*. URL: https://www.nltk.org/nltk_data/.

[7] *ML CO_2 impact*. URL: https://mlco2.github.io/impact/#compute.

[8] *Plain text Wikipedia (SimpleEnglish)*. Accessed: April 2024. kaggle. URL: https://www.kaggle.com/datasets/ffatty/plain-text-wikipedia-simpleenglish.

[9] Victor Sanh et al. *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. 2020. arXiv: 1910.01108 [cs.CL].