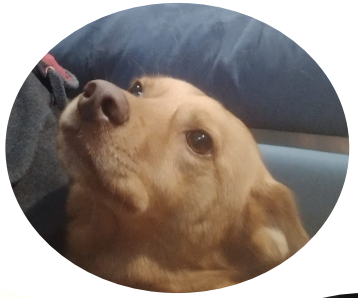
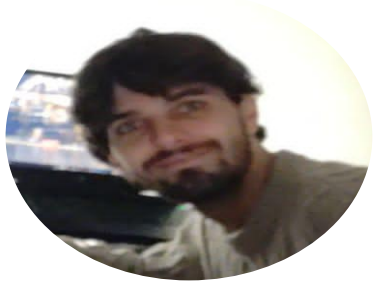
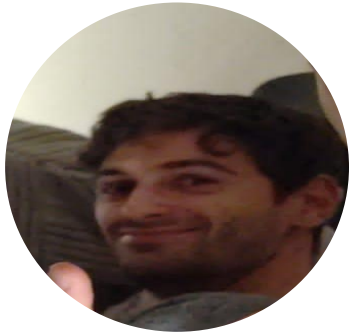

PROJECT
ATHENA

Who we are

- Giordano Scerra
- Andrea Marino
- Yuri Negri
- Davide Borghini
- Davide Marchi


Presentation Outline

Introduction → Data analysis → Metrics

Hypotheses → Models → Results & conclusions

Introduction

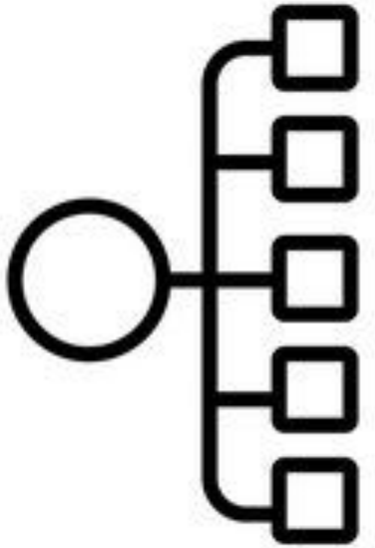
Introduction

- **Multinomial** text classification
- **Multidisciplinary** approach
- Dataset **exploration**
- Texts from **Project Gutenberg**
- A few **hypotheses**, different **models**

Goal of the project

The goal of the project:

- **Classify** the philosophical current of sentences in the dataset, using different models
- **Compare** the models' performance
- **Assess** our hypotheses concerning the data composition



Previous results

The Philosophy Data Project
classifiers:

- Naive Bayes 77% accuracy (on validation)
- Recurrent models 80% accuracy (on validation)

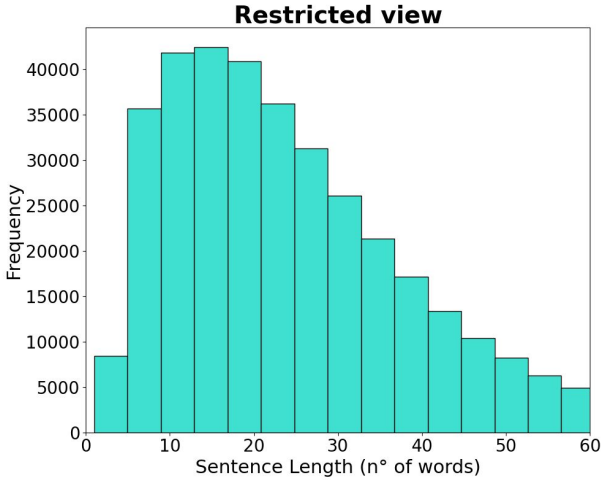
Kourosh Alizadeh

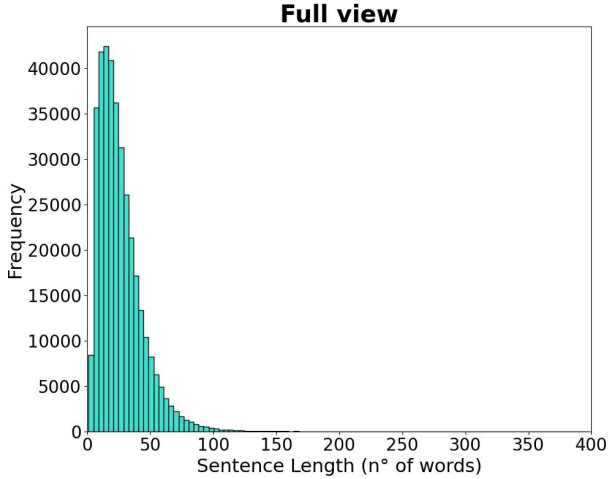
Data Analysis

The Dataset

From **Kaggle**:

- 59 books
- 36 authors
- 13 philosophical currents
- 360.808 sentences

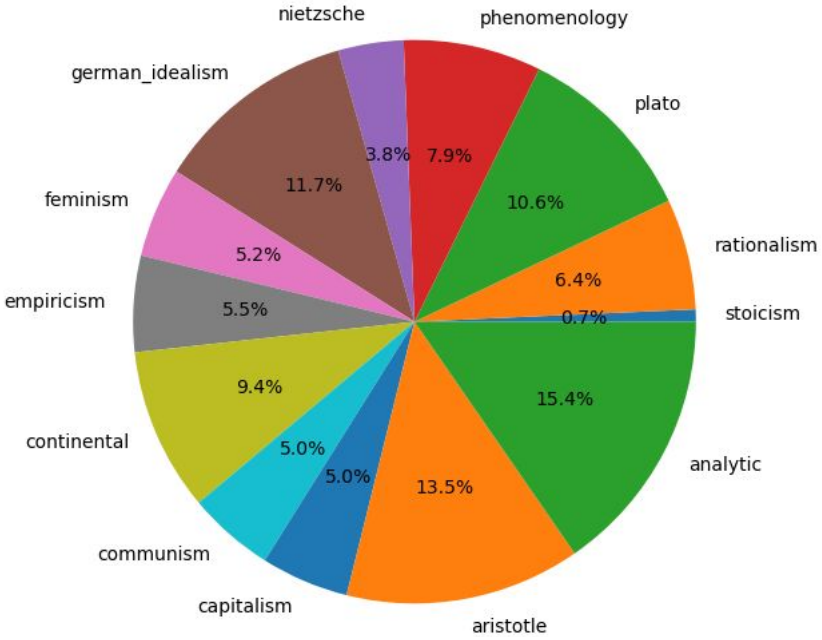




Philosophical currents

The **schools** in the dataset are:

- Plato
- Aristotle
- Empiricism
- Rationalism
- Analytic
- Continental
- Phenomenology
- German Idealism
- Communism
- Capitalism
- Stoicism
- Nietzsche
- Feminism



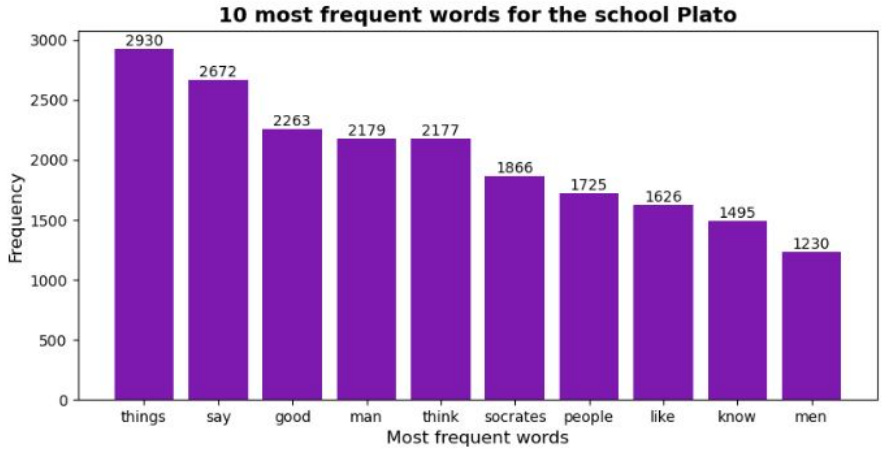
Their distribution is quite **skewed**

Data analysis

Histograms

Histograms with the count of the 10 most frequent words for each class have been created, such as the one on the right.

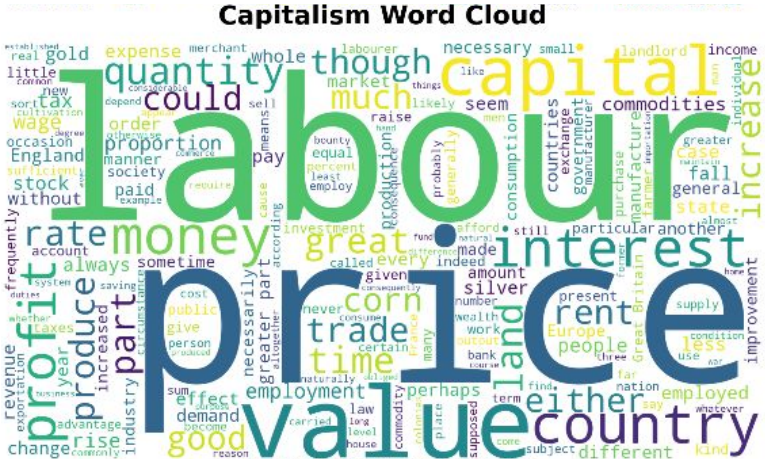
For each class, few words make up the majority of the distribution, as it can be seen also in the wordclouds.



Data analysis

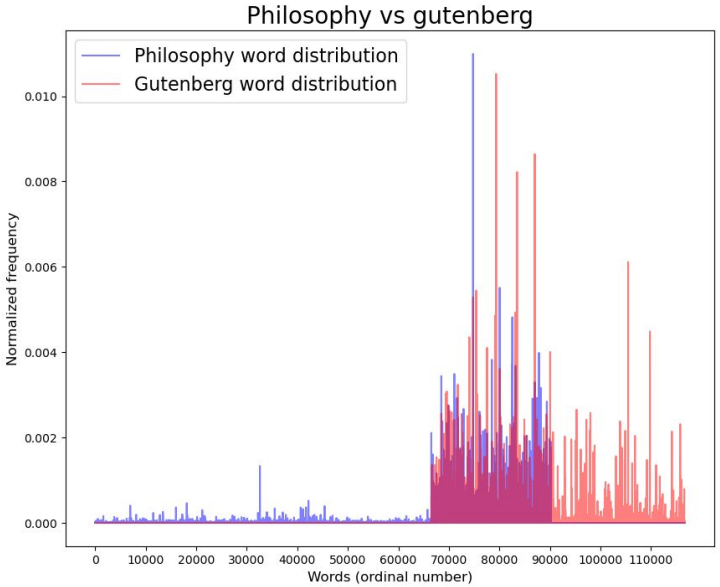
Word clouds

Classes parallelism through the word *“labour”*



Contrastive Analysis

Word Usage Distribution: Gutenberg corpus



Gutenberg corpus: already available in NLTK, and intuitively similar to the philosophy one.

90.92% of the words in the philosophy corpus are shared with the gutenberg corpus

88.86% of the words in the Gutenberg corpus are shared with our philosophy corpus

The distance of the two distributions (on the shared portion of the support) is 2.8719e-02

The cosine similarity between the two distributions (on the shared portion of the support) is 0.5670

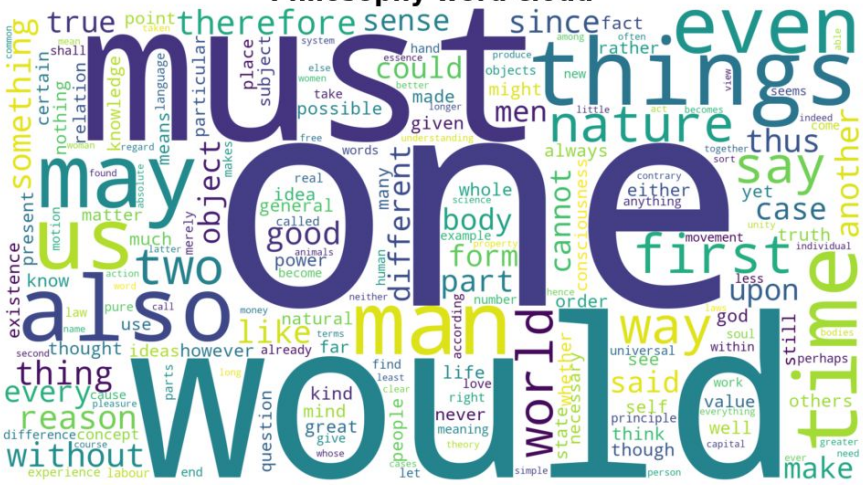
Contrastive Analysis

Word Usage Distribution: Gutenberg corpus

Word clouds for the Gutenberg and Philosophy corpora

Gutenberg word cloud

Philosophy word cloud



Preprocessing

To **convert the text** into a processable format:

- Pdf converted to txt
- Heavy correction
- Spacy for tokenization
- Drop short phrases
- Further cleaning
- Dataframe creation

Metrics

Metrics: philosopher choice

Considering the approaches of the author we already have a **baseline result** to start from.

An accurate analysis of his work gave us ideas regarding **untried approaches**, which we explored during the course.

Contrary to his work, we chose F1 score macro average as a reference metric.

Metrics: our choice

Motivations:

- Unbalanced dataset
- Classes distribution is not to be considered representative of the real word.
- False positive and false negative should be equally weighted.

Hypotheses

Hypotheses

Sentence length and Meaning

Many short sentences in the dataset could come from different schools and some aren’t well segmented.

Our hypothesis are that they act akin to **stopwords** and a portion of them is **noise**.

We decided to train and validate on both the full dataset and a reduced version of it.

The reduced version is trimmed of all the sentences that are composed by < **85 characters**

Hypotheses

Formalization

What we expected:

- **α>□ and γ>δ**: shorter sentences are inherently hard to be classified by any model, whether it sees all TR data or not.
- **α>γ and □>δ**: short sentences act as noise from which no meaningful regularity can be learned. Thus, they degrade the performance of any model.

Table: proportion of TR (rows), proportion of VL (columns)

	≥ 84	Full
≥ 84	α	β
Full	γ	δ

Models

Evaluated models

We evaluated the following models:

- Naive Bayes
- Recurrent Neural Networks
- BERT
- DistilBert
- Zero-shot bart-large-mnli

On these splits:

- 90% of data into the Devset
  - 20% Of these into the Validation set
- 10% of data into the Test set
- Stratified w.r.t the schools

Naive Bayes

Approaches

We tried some variations:

- No stopwords
- Negation token
- tf-idf
- Count vectorizer
- Grid search on smoothing parameter

Recurrent Neural Networks

Architectures

For the RNNs, we tried different architectures

- Pretrained GloVe vectors of dimension 100 and 200
- Untrained embeddings of dimension 128
- Bidirectional layers of 32 (64) units
  - Gated Recurrent Units
  - Long Short Term Memory Units
  - Simple RNN cells

What worked best: untrained embeddings and bidirectional LSTM, with Adam optimizer

2 Berts with one stone

The best model

BERT by Google is the largest transformer-based model we have fine-tuned.

We used all phrases except for one that was too long.

We fine-tuned the BERT model in 3 epochs using the pre-defined model ***'BertForSequenceClassification'*** (no need to reinvent the wheel).

DistilBert

Baby Bert

HEAD:
- relu
- dropout
- sigmoid

Same as BERT but smaller. Trained by distillation of the pretrained BERT model, meaning it's been trained to predict the same probabilities as the larger model. The actual objective is a combination of:

- finding the same probabilities as the teacher model
- predicting the masked tokens correctly (but no next-sentence objective)
- a cosine similarity between the hidden states of the student and the teacher model

Zero Shot Learning

And why it didn’t work

We used the BART-large-MNLI model by Facebook, the largest model we tried.

We tested different labels:

- Unmodified: ***analytic***, *aristotle, rationalism, continental…*
- Descriptive: *Analytic Philosophy,* ***Continental Philosophy****…*

The results were underwhelming. Due to the large number of closely related classes, this non-fine-tuned approach tended to skew towards more generic and "umbrella" options.

Environmental impact

Not so small

Approximately 29 hours of GPU usage

- 15 h Bert
- 8 h DistilBert
- 6 h Zero shot

This led up to 3,76 kg of CO_2 equivalent

- 1,63L of fuel burned
- 140g of cow meat produced

Results & Conclusions

Results

Full and Reduced

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

The model that performed best was BERT trained over the full development set. Hence, we tested it over the test sets:

- **Full**: 0.83
- **≥ 84**: 0.88

Conclusions

- The classifiers we built can't actually tell to which philosophical current a sentence belongs to.
    - "only" able to classify the topic of a phrase.
- Complex models are only slightly better than simpler models. Trade-off!
- Improved results of Kourosh Alizadeh by a good margin, but at what cost?
- Our hypothesis was slightly wrong.

Conclusions

On our hypothesis

Table: proportion of TR (rows), proportion of VL (columns)

	≥ 84	Full
≥ 84	α	β
Full	γ	δ

α>□ and γ>δ were always true (aside from zero-shot). Then, short sentences are actually difficult to classify, we assume it's because most of them bring little meaning.

α>γ and □>δ were always false. Then, shorter sentences do not act as noise during training, but instead they help the model to classify also longer sentences. We suppose this happens because we removed all the short sentences, though some can convey meaning.

“I guess there are never enough books.”

— John Steinbeck

THANKS FOR THE ATTENTION

Giordano Scerra

Andrea Marino

Yuri Negri

Davide Marchi

Davide Borghini