Human Values Behind Arguments

(c) Sebastian Pellizzari, Peter Burger, Elbaraa Elsaadany, Hamed Homaeirad and Jannis David Voigt

The Big Picture

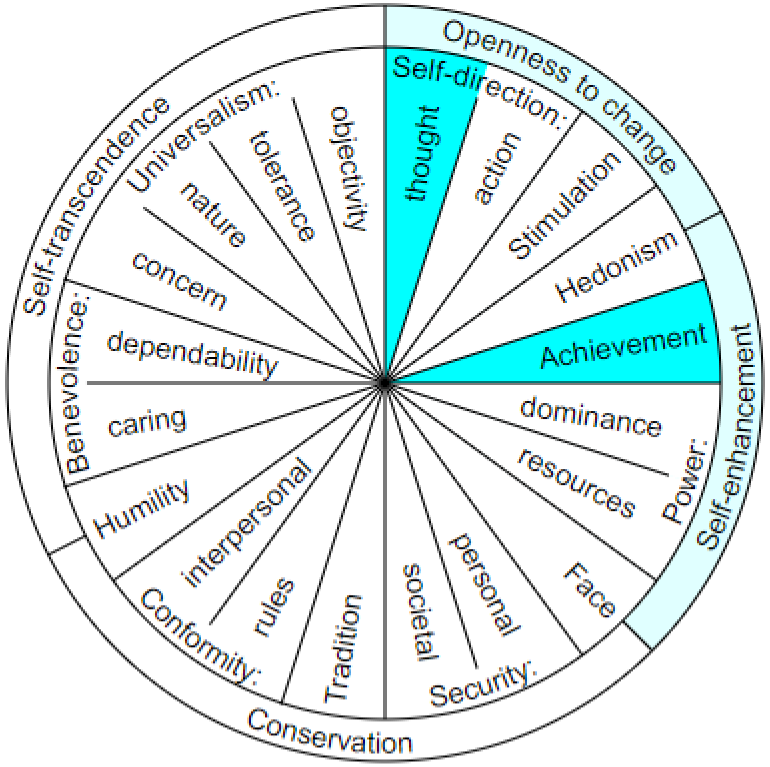
- Task: Any prompt should be automatically labeled with 1+ labels

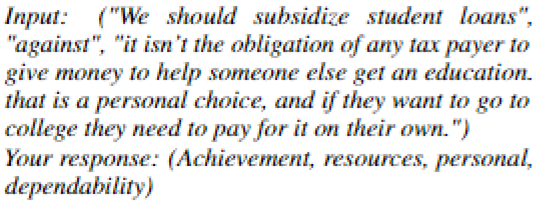
  Multi-label Classification Problem

- Prompts/Inputs consist of triplets:

  <premise>, <stance>, <justification>

- Hierarchy of Label Levels:
  - 4 levels (original dataset, 2021)
  - 2 levels (revised dataset, 2023)
  - Sub-levels do not necessarily fall under one fixed upper level (e.g., Humility, Face, Hedonism)

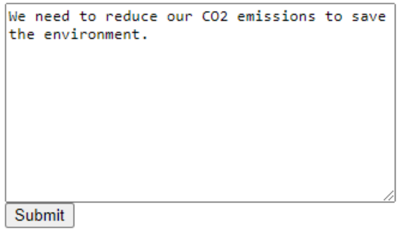




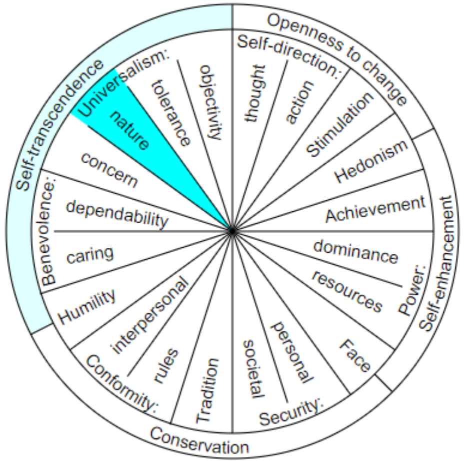
Input: ("We should subsidize student loans", "against", "it isn't the obligation of any tax payer to give money to help someone else get an education. that is a personal choice, and if they want to go to college they need to pay for it on their own.")
Your response: (Achievement, resources, personal, dependability)

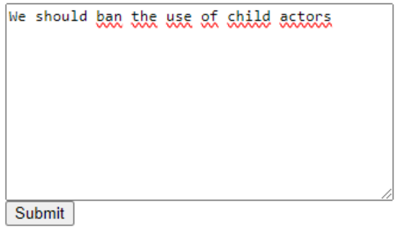
https://values.args.me/
(last acessed: 23.01.2024)

The Big Picture

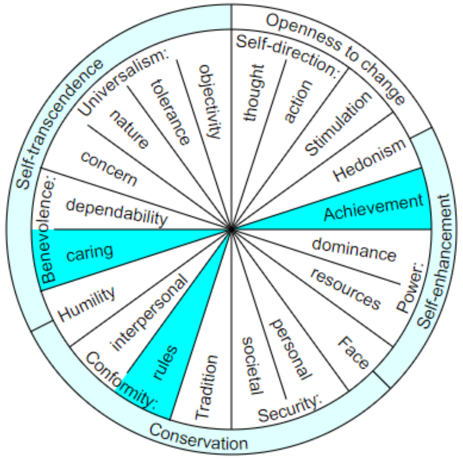


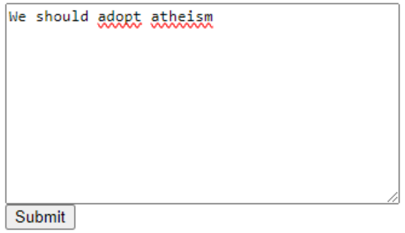
We need to reduce our CO2 emissions to save the environment.



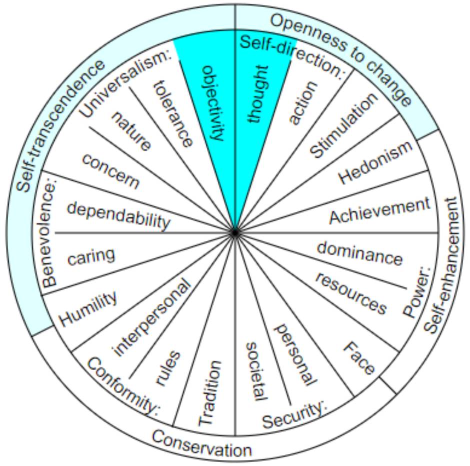


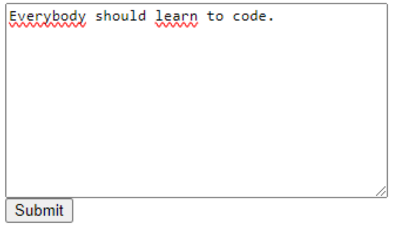
We should ban the use of child actors



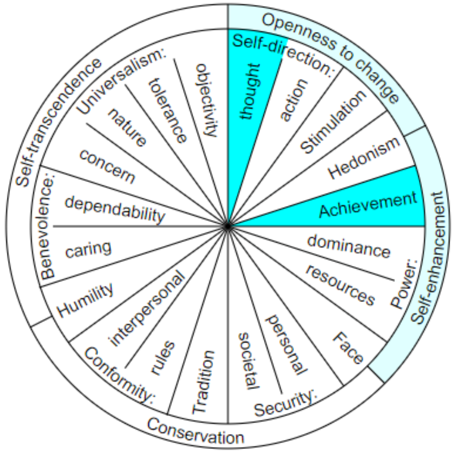


We should adopt atheism





Everybody should learn to code.



https://values.args.me/

(All from, last acessed: 23.01.2024)

An Overview – What Has Been Done?

- Research Questions
- Exploratory Data Analysis (EDA)
- Experiments on ChatGPT3.5
- Experiments on Llama-2-7b

The Research Questions

- Are there any inconsistencies in the dataset?
- Does the presence of any label ‘imply’ another label?
- How well the general LLMs compare to the fine-tuned LMs (such as BERT, etc.) participating in SemEval-2023 competition?
- Will the result be any different if role-based prompts are used? (e.g., prompts asked from the perspective of a scientist or a psychologist)

Exploratory Data Analysis (EDA) – What is that?

Formal Definition: *Exploratory data analysis […] is used by data scientists to analyze […] data sets and summarize their main characteristics […]. It helps determine how best to manipulate data sources to get the answers you need, making it easier […] to discover patterns, spot anomalies, test a hypothesis, or check assumptions.* *(https://www.ibm.com/topics/exploratory-data-analysis, Last access: 21.01.2024)*

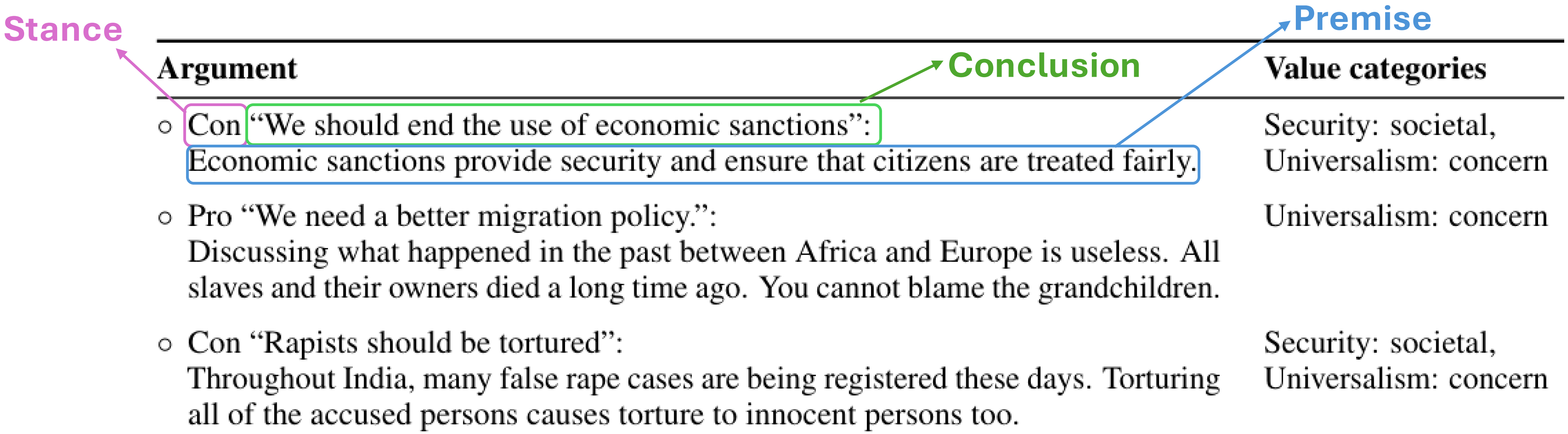
Initial Data Insights

# of	Data Split			
	Train	**Validation**	**Test**	**Total**
samples	5393 (60.83%)	1896 (21.39%)	1576 (17.78%)	8865 (100%)
missing values (initial assessment)	0	0	0	0

Label assignments for samples are done via a one-hot encoding like manner.
Therefore, it should also be checked **if all the samples are assigned to at least one label**.

Touché23 Dataset
Version 2023-08-15
https://zenodo.org/records/8248658

Feature Analysis – General Facts



Stance

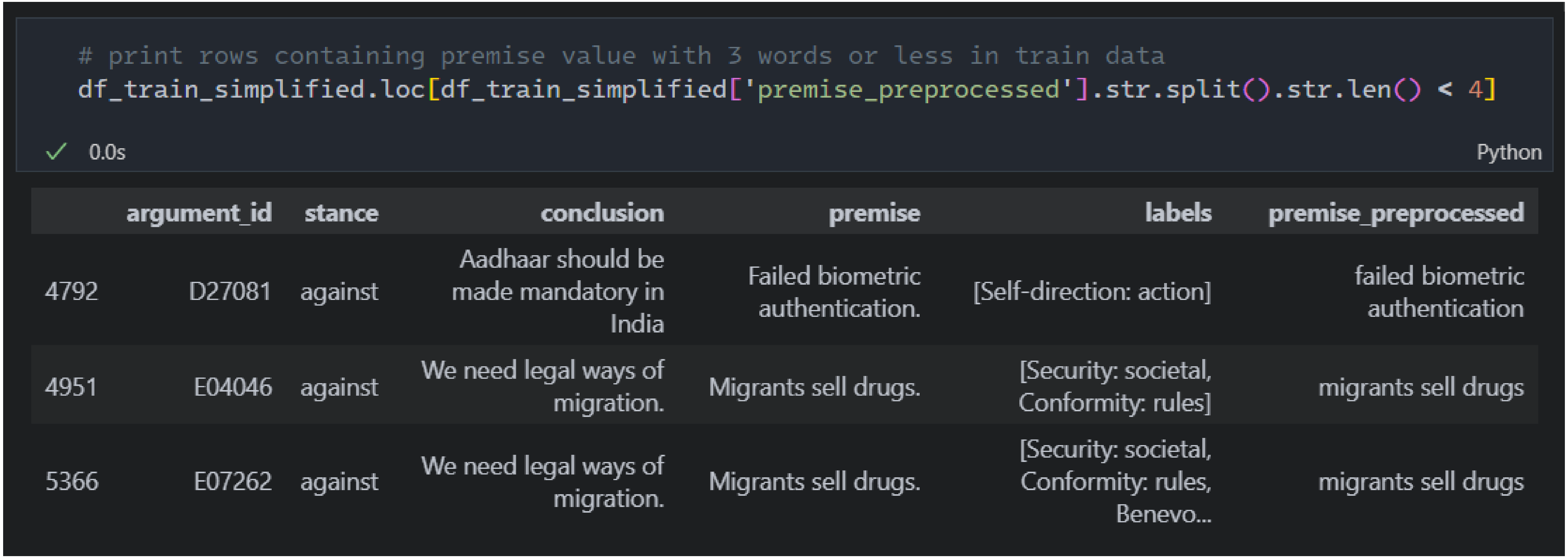
Conclusion

Premise

Argument	Value categories
○ Con “We should end the use of economic sanctions”: Economic sanctions provide security and ensure that citizens are treated fairly.	Security: societal, Universalism: concern
○ Pro “We need a better migration policy.”: Discussing what happened in the past between Africa and Europe is useless. All slaves and their owners died a long time ago. You cannot blame the grandchildren.	Universalism: concern
○ Con “Rapists should be tortured”: Throughout India, many false rape cases are being registered these days. Torturing all of the accused persons causes torture to innocent persons too.	Security: societal, Universalism: concern

Statistics Val.	"Premise" Length	
	# of words	# of characters
Min.	3	19
Max.	133	780
Avg.	21	126

Feature Analysis – Digging Deeper



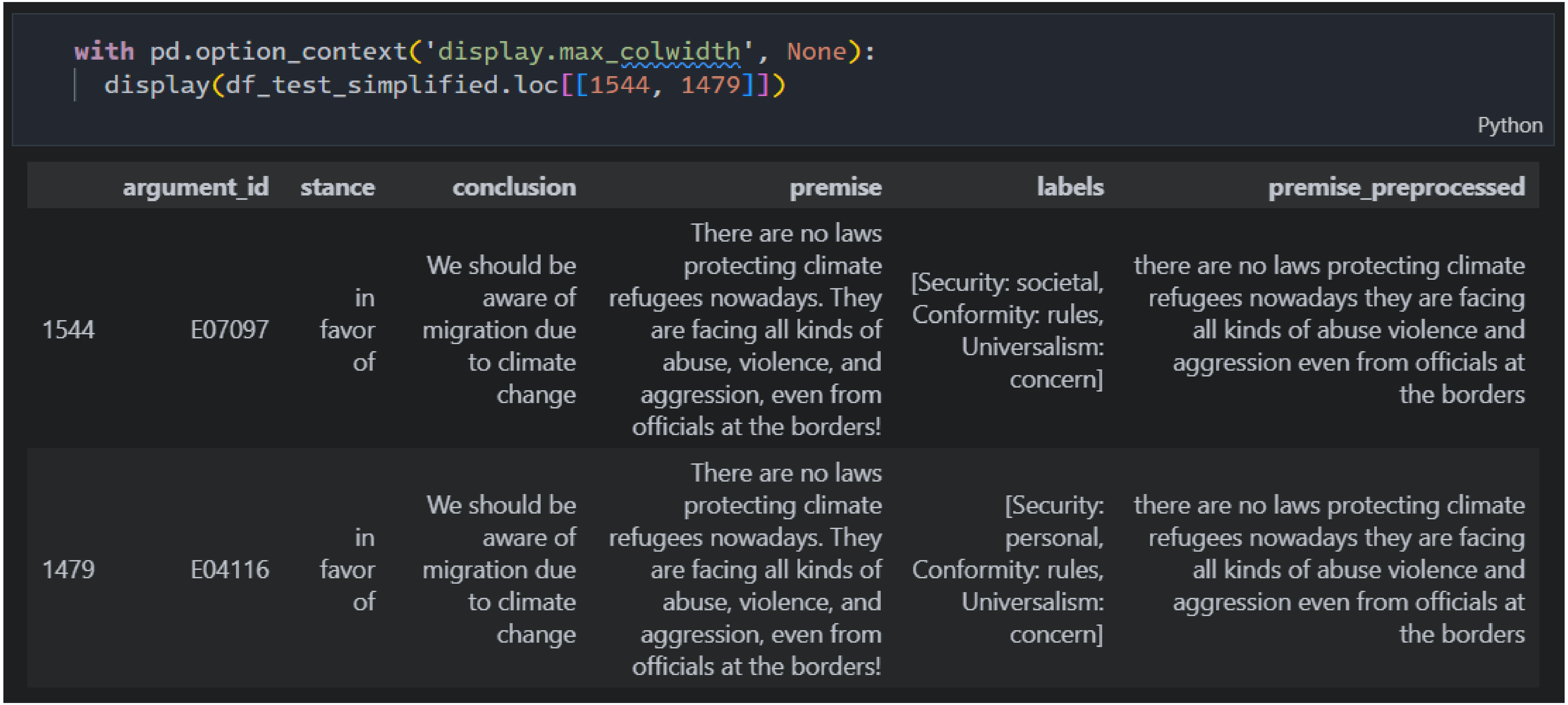
```python
# print rows containing premise value with 3 words or less in train data
df_train_simplified.loc[df_train_simplified['premise_preprocessed'].str.split().str.len() < 4]
```

✓ 0.0s

Python

	argument_id	stance	conclusion	premise	labels	premise_preprocessed
4792	D27081	against	Aadhaar should be made mandatory in India	Failed biometric authentication.	[Self-direction: action]	failed biometric authentication
4951	E04046	against	We need legal ways of migration.	Migrants sell drugs.	[Security: societal, Conformity: rules]	migrants sell drugs
5366	E07262	against	We need legal ways of migration.	Migrants sell drugs.	[Security: societal, Conformity: rules, Benevo...	migrants sell drugs

Not Just Train Data...

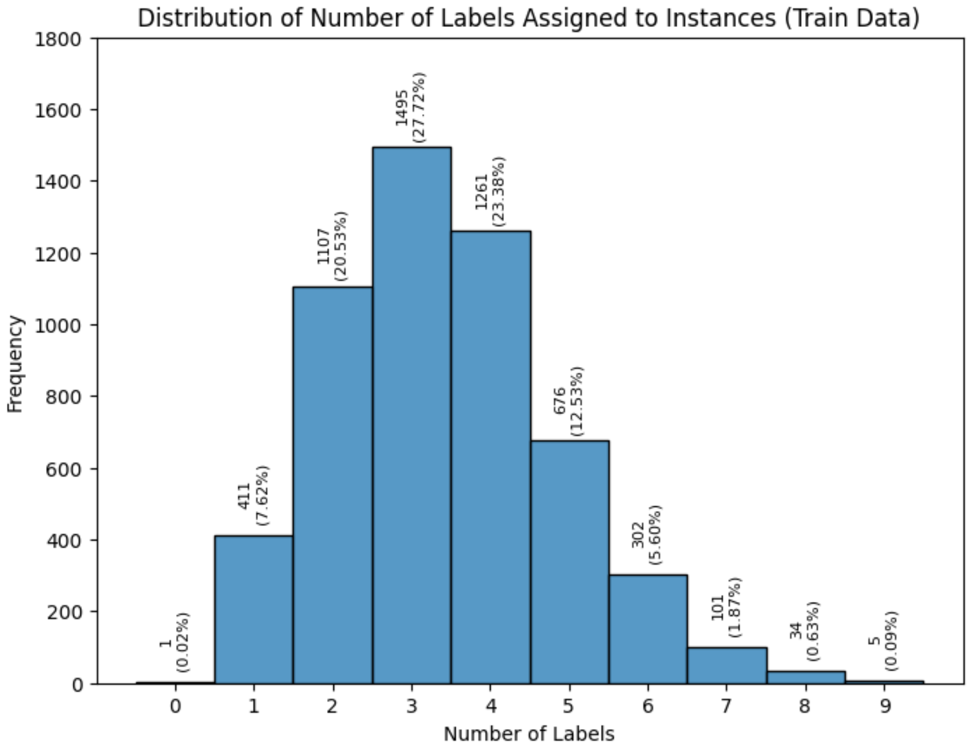


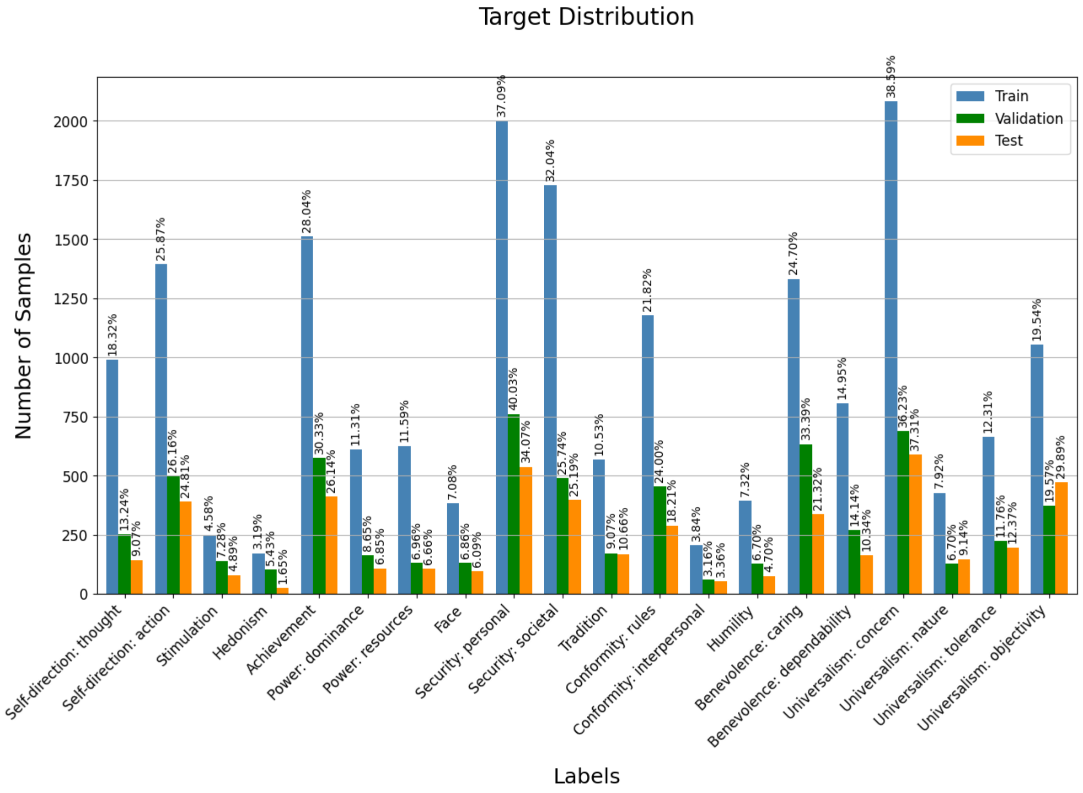
```python
with pd.option_context('display.max_colwidth', None):
    display(df_test_simplified.loc[[1544, 1479]])
```

	argument_id	stance	conclusion	premise	labels	premise_preprocessed
1544	E07097	in favor of	We should be aware of migration due to climate change	There are no laws protecting climate refugees nowadays. They are facing all kinds of abuse, violence, and aggression, even from officials at the borders!	[Security: societal, Conformity: rules, Universalism: concern]	there are no laws protecting climate refugees nowadays they are facing all kinds of abuse violence and aggression even from officials at the borders
1479	E04116	in favor of	We should be aware of migration due to climate change	There are no laws protecting climate refugees nowadays. They are facing all kinds of abuse, violence, and aggression, even from officials at the borders!	[Security: personal, Conformity: rules, Universalism: concern]	there are no laws protecting climate refugees nowadays they are facing all kinds of abuse violence and aggression even from officials at the borders

Target Analysis:

Distribution of Labels in Data





- Average number of labels: 3
- Minimum number of labels: **0 (?)**
- Maximum number of labels: 9

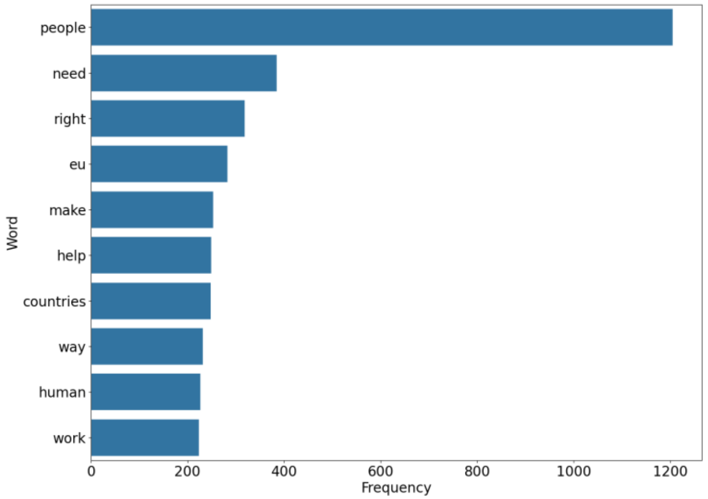
(out of 20 labels)

Ratio of labelled samples don't follow the same distribution for training, validation, and test splits.
→ Any training or fine-tuning is prone to resulting in biased models!

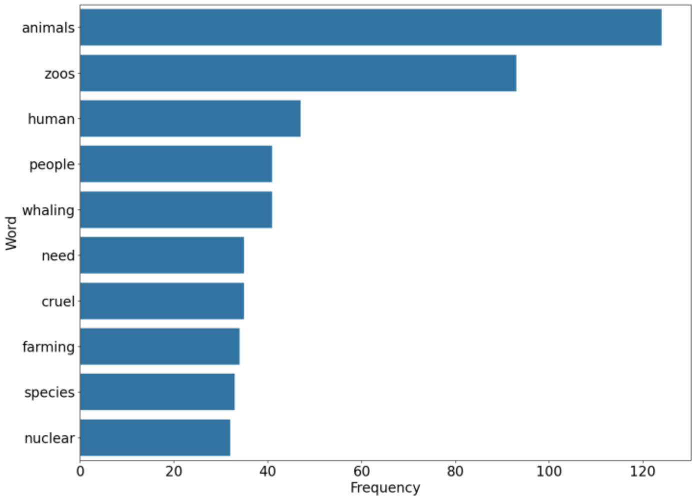
Do certain words in Premise imply a certain label?

Frequency of word appearance in "Premise" column of Train Data

(stop-words are ignored: i.e., is, the, a, ...)

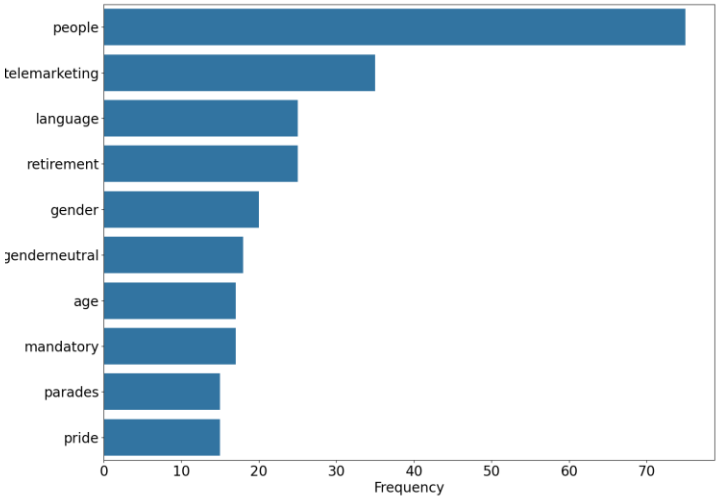


Most used words in dataset for all instances



Most used words in dataset with label:

Universalism – nature

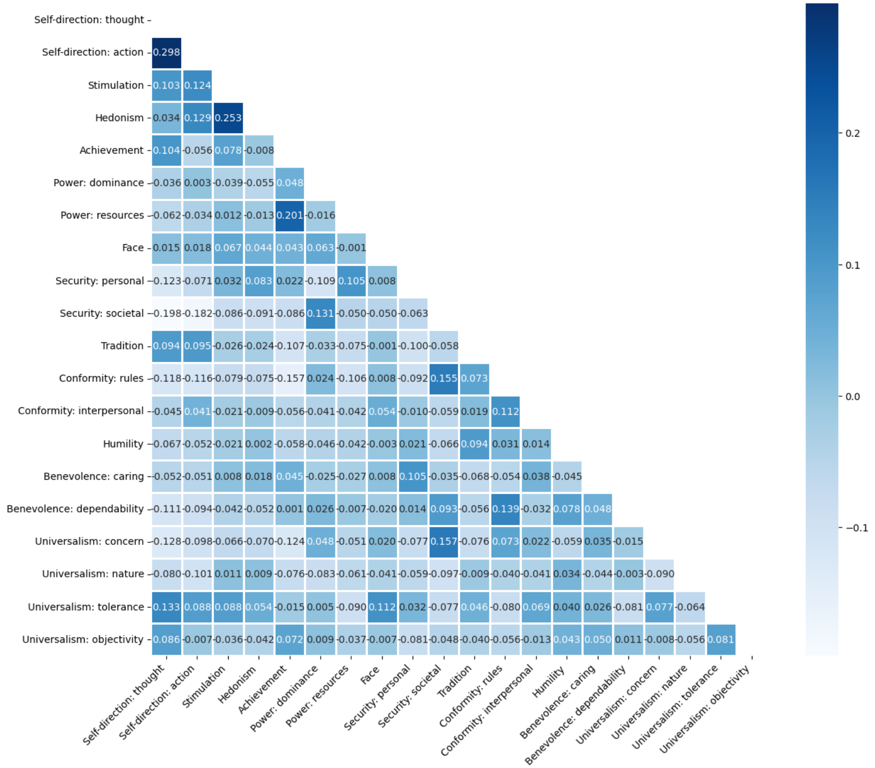


Most used words in dataset with label:

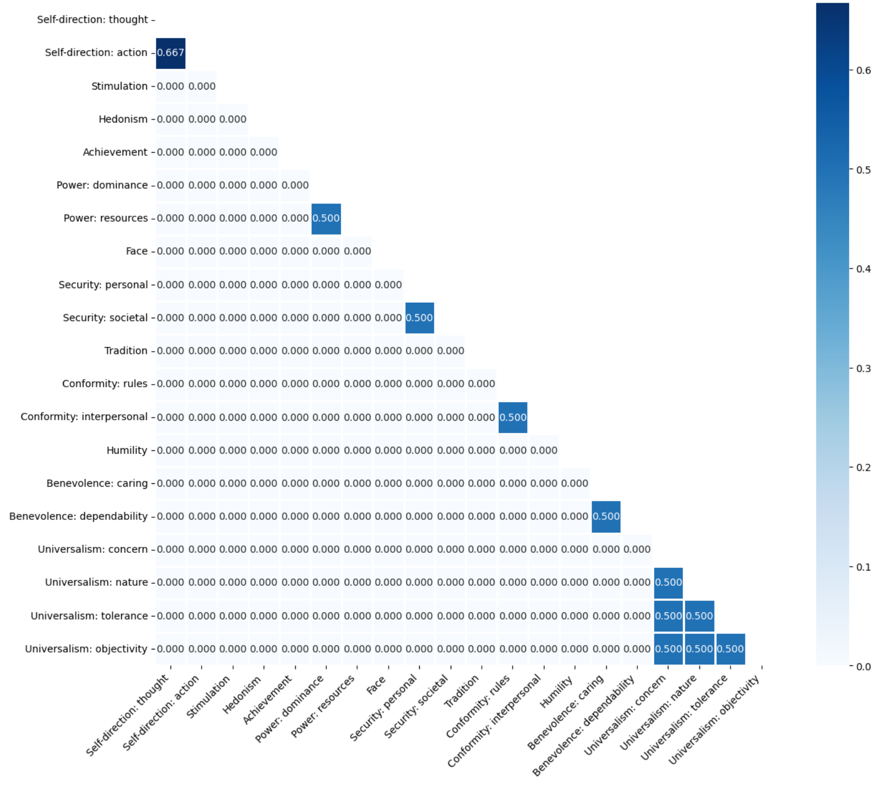
Conformity: inter-personal

Correlation between labels?

...hardly



	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
Self-direction: thought																				
Self-direction: action	0.298																			
Stimulation	0.103	0.124																		
Hedonism	0.034	0.129	0.253																	
Achievement	0.104	-0.056	0.078	-0.008																
Power: dominance	-0.036	0.003	-0.039	-0.055	0.048															
Power: resources	-0.062	-0.034	0.012	-0.013	0.201	-0.016														
Face	0.015	0.018	0.067	0.044	0.043	0.063	-0.001													
Security: personal	-0.123	-0.071	0.032	0.083	0.022	-0.109	0.105	0.008												
Security: societal	-0.198	-0.182	-0.086	-0.091	-0.086	0.131	-0.050	-0.050	-0.063											
Tradition	0.094	0.095	-0.026	-0.024	-0.107	-0.033	-0.075	-0.001	-0.100	-0.058										
Conformity: rules	-0.118	-0.116	-0.079	-0.075	-0.157	0.024	-0.106	0.008	-0.092	0.155	0.073									
Conformity: interpersonal	-0.045	0.041	-0.021	-0.009	-0.056	-0.041	-0.042	0.054	-0.010	-0.059	0.019	0.112								
Humility	-0.067	-0.052	-0.021	0.002	-0.058	-0.046	-0.042	-0.003	0.021	-0.066	0.094	0.031	0.014							
Benevolence: caring	-0.052	-0.051	0.008	0.018	0.045	-0.025	-0.027	0.008	0.105	-0.035	-0.068	-0.054	0.038	-0.045						
Benevolence: dependability	-0.111	-0.094	-0.042	-0.052	0.001	0.026	-0.007	-0.020	0.014	0.093	-0.056	0.139	-0.032	0.078	0.048					
Universalism: concern	-0.128	-0.098	-0.066	-0.070	-0.124	0.048	-0.051	0.020	-0.077	0.157	-0.076	0.073	0.022	-0.059	0.035	-0.015				
Universalism: nature	-0.080	-0.101	0.011	0.009	-0.076	-0.083	-0.061	-0.041	-0.059	-0.097	-0.009	-0.040	-0.041	0.034	-0.044	-0.003	-0.090			
Universalism: tolerance	0.133	0.088	0.088	0.054	-0.015	0.005	-0.090	0.112	0.032	-0.077	0.046	-0.080	0.069	0.040	0.026	-0.081	0.077	-0.064		
Universalism: objectivity	0.086	-0.007	-0.036	-0.042	0.072	0.009	-0.037	-0.007	-0.081	-0.048	-0.040	-0.056	-0.013	0.043	0.050	0.011	-0.008	-0.056	0.081	



	Self-direction: thought	Self-direction: action	Stimulation	Hedonism	Achievement	Power: dominance	Power: resources	Face	Security: personal	Security: societal	Tradition	Conformity: rules	Conformity: interpersonal	Humility	Benevolence: caring	Benevolence: dependability	Universalism: concern	Universalism: nature	Universalism: tolerance	Universalism: objectivity
Self-direction: thought																				
Self-direction: action	0.667																			
Stimulation	0.000	0.000																		
Hedonism	0.000	0.000	0.000																	
Achievement	0.000	0.000	0.000	0.000																
Power: dominance	0.000	0.000	0.000	0.000	0.000															
Power: resources	0.000	0.000	0.000	0.000	0.000	0.500														
Face	0.000	0.000	0.000	0.000	0.000	0.000	0.000													
Security: personal	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000												
Security: societal	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.500											
Tradition	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000										
Conformity: rules	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000									
Conformity: interpersonal	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.500								
Humility	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000							
Benevolence: caring	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000						
Benevolence: dependability	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.500					
Universalism: concern	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000				
Universalism: nature	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.500			
Universalism: tolerance	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.500		
Universalism: objectivity	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.500	0.500	

- **Left: Bi-variant Analysis; Right: Cosine Similarity**
- **Highest correlation between labels is “Self-direction: thought” and “Self-direction action” with correlation of 0.3 (=fairly low)**
- **However: Correlation ≠ Causation !**

Summary of Inconsistencies in Dataset

Missing Labels & Duplicate Inputs

	Train	Validation	Test	Total
# of missing labels	**1**	**0**	**0**	**1**
# of duplicate values	**89**	**35**	**32**	**156**

Target Label Distribution

Ratio of labeled samples per class in preset splits of Train, Validation & Test data do not necessarily follow the same distribution, sometimes off by a noticeable margin.

Prompt - Techniques

- Role setting
- Zero-shot
- Few-shot
- Chain-of-thought
- YAML for structured input, output, context

Prompt - Insight

Imagine you are a scientist. It is your job to analyze the human values behind a list of arguments ...

An argument is thereby given as a triple (premise, stance, conclusion). The premise is the statement given to the person. The stance is ...

Use following values:

- *thought: it is good to have own ideas and interests*
- *action: it is good to determine one's own actions ...*

Examples:

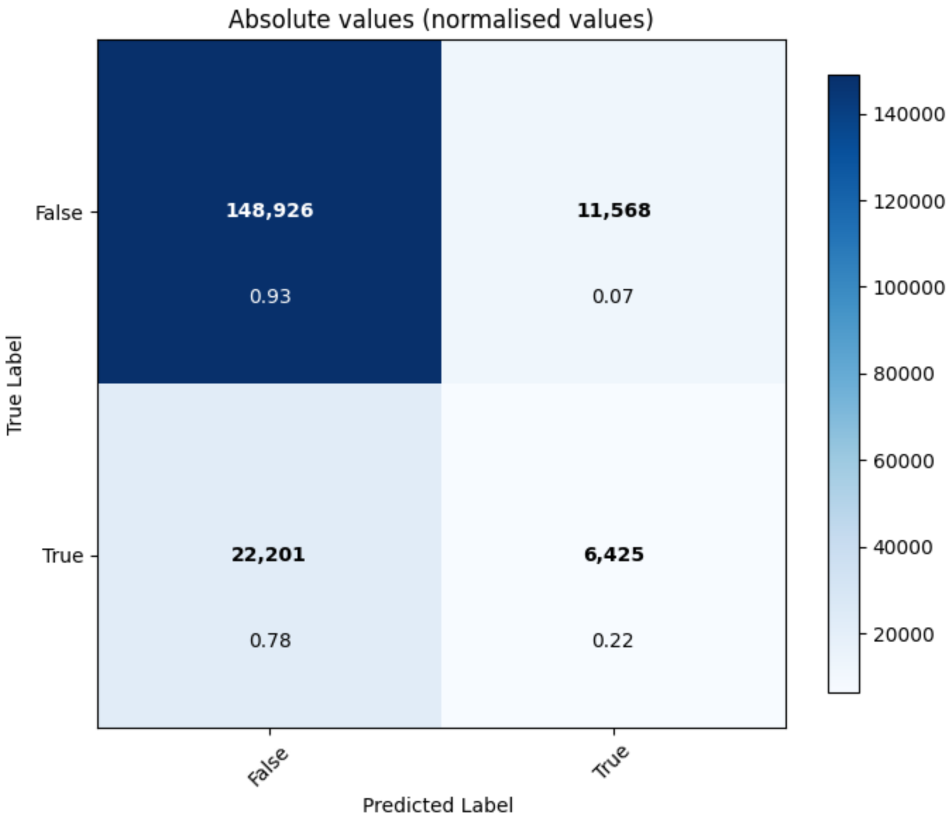
Input: ("We should cancel pride parades","in favor of","pride parades create a huge disturbance"),

Output: (rules, interpersonal)...

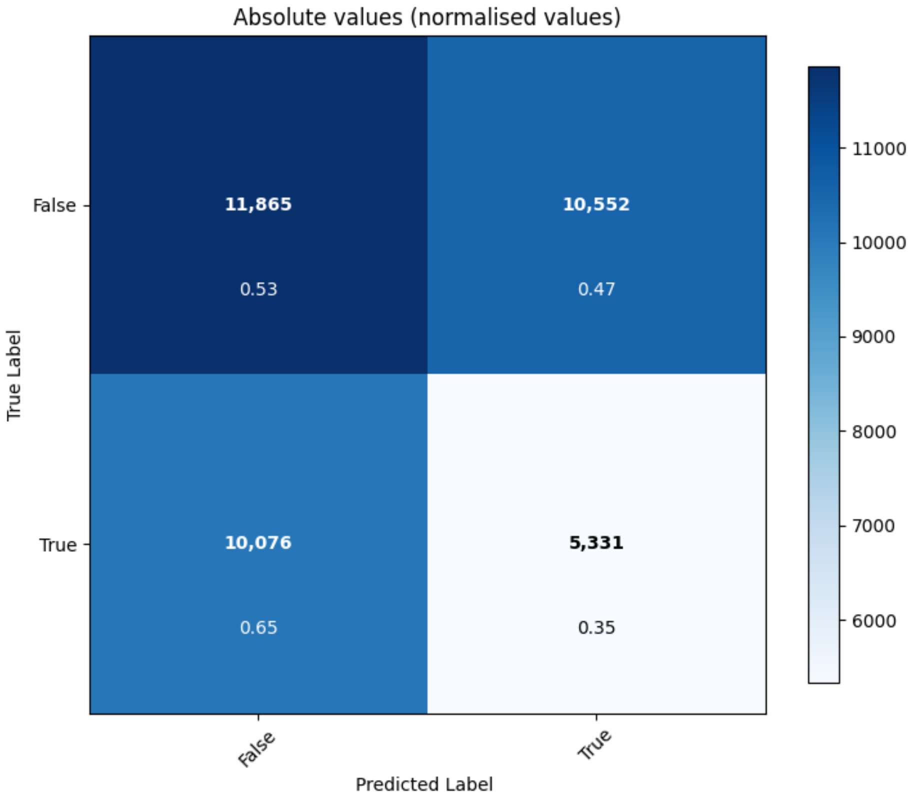
Input: ("We should end mandatory retirement","in favor of","mandatory retirement is purely age discrimination")

Output:

ChatGPT - Results

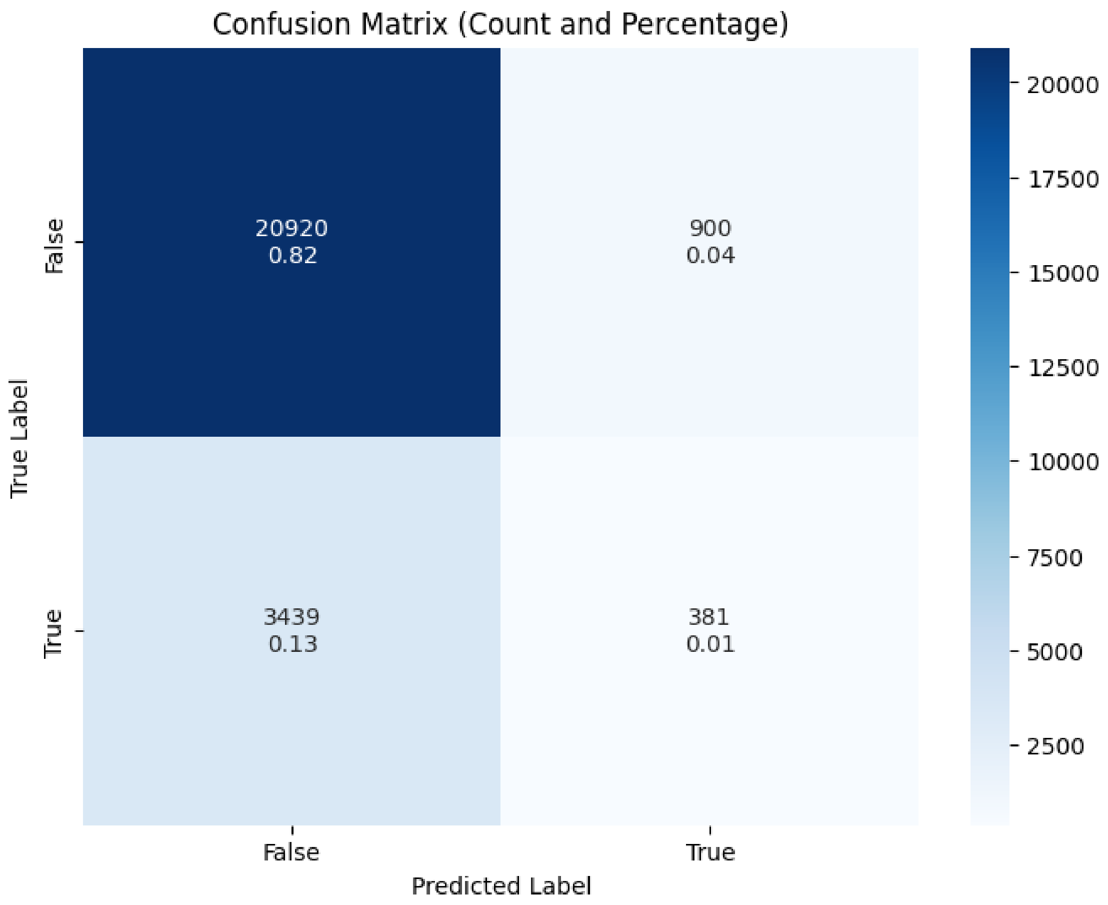


Confusion matrix fine granularity (F1 ~ 0.275553)



Confusion coarse granularity (F1 ~ 0.79924)

Llama2- Results



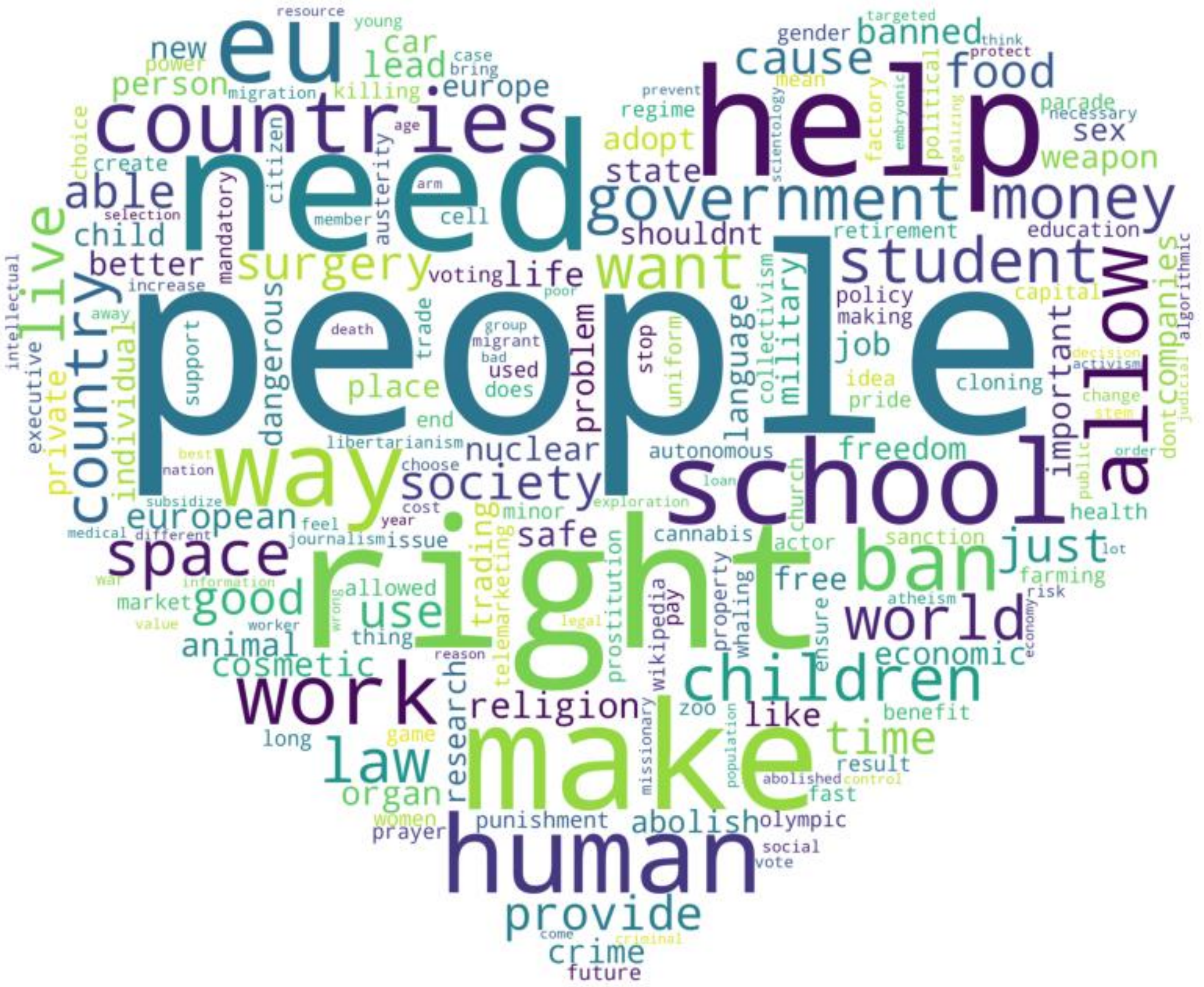
Confusion coarse granularity (F1 ~ 0.15)

Conclusions

- Developers of ML-based solutions for problems should prioritize thorough analysis of the data
  - EDA is a MUST

- Smaller language models prior to the LLMS are still capable if fine-tuned for a specific task
  - This makes them ideal for utilizing them into Production pipelines (faster inference time, less costly)

QA



Word Frequency Cloud Generated from "Premise" Column of Training Data