Beyond Labels

Jennifer Caceres, Jenifer Vivar, Ali Salem, Zoya Shafique

Agenda

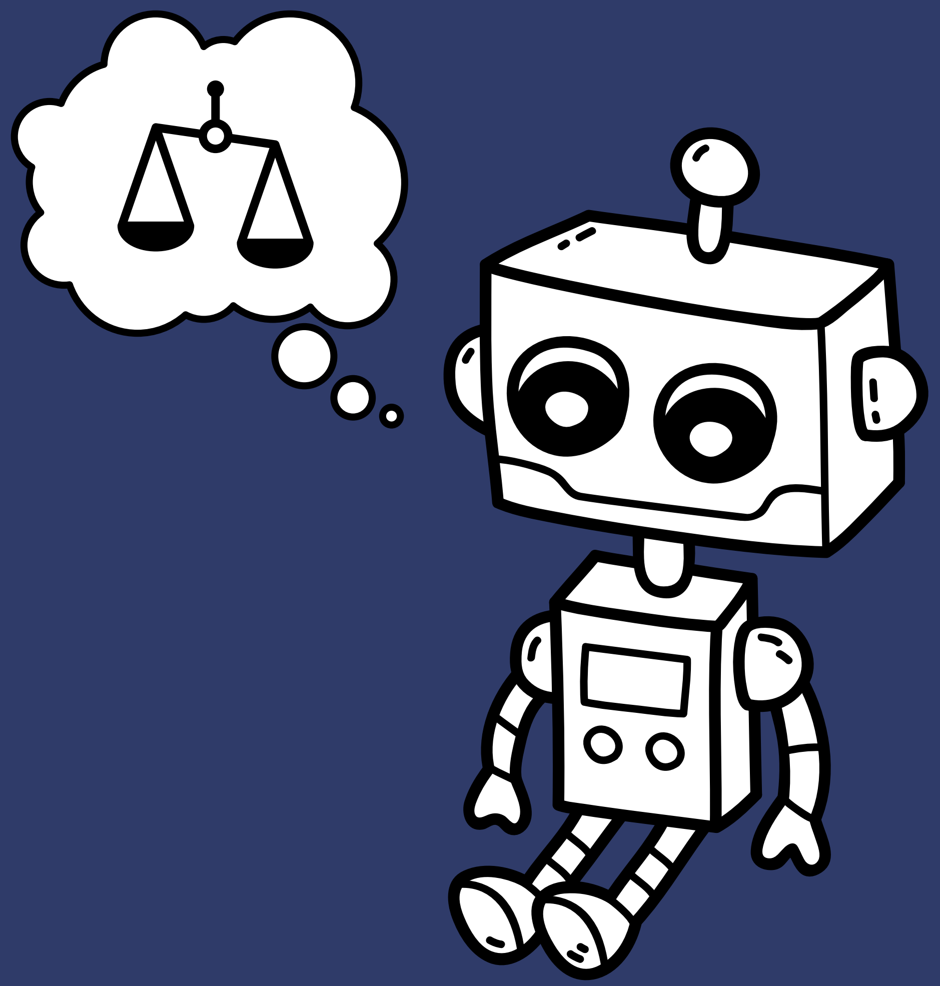
1. Introduction
2. What is Moral Foundational Theory?
3. Datasets
4. Big Picture Roadmap
5. Unsupervised Baselines
6. Semi-supervised Approach
8. Conclusions

Introduction: Problem Statement

- Given a piece of text, can we find the underlying moral views ?
- Can we train models with reduced supervision to classify morals in text?

Introduction: Motivations

With the rise of large language models and AI such as ChatGPT, it is of increasing importance to preserve moral and ethical foundations in AI applications.



What is Moral Foundational Theory?

- It aims to explain the origins and variations of moral judgments and values across different cultures and individuals.
- Main groups of moral foundations

Care/Harm

Loyalty/Betrayal

Liberty/Oppression

Fairness/Cheating

Authority/Subversion

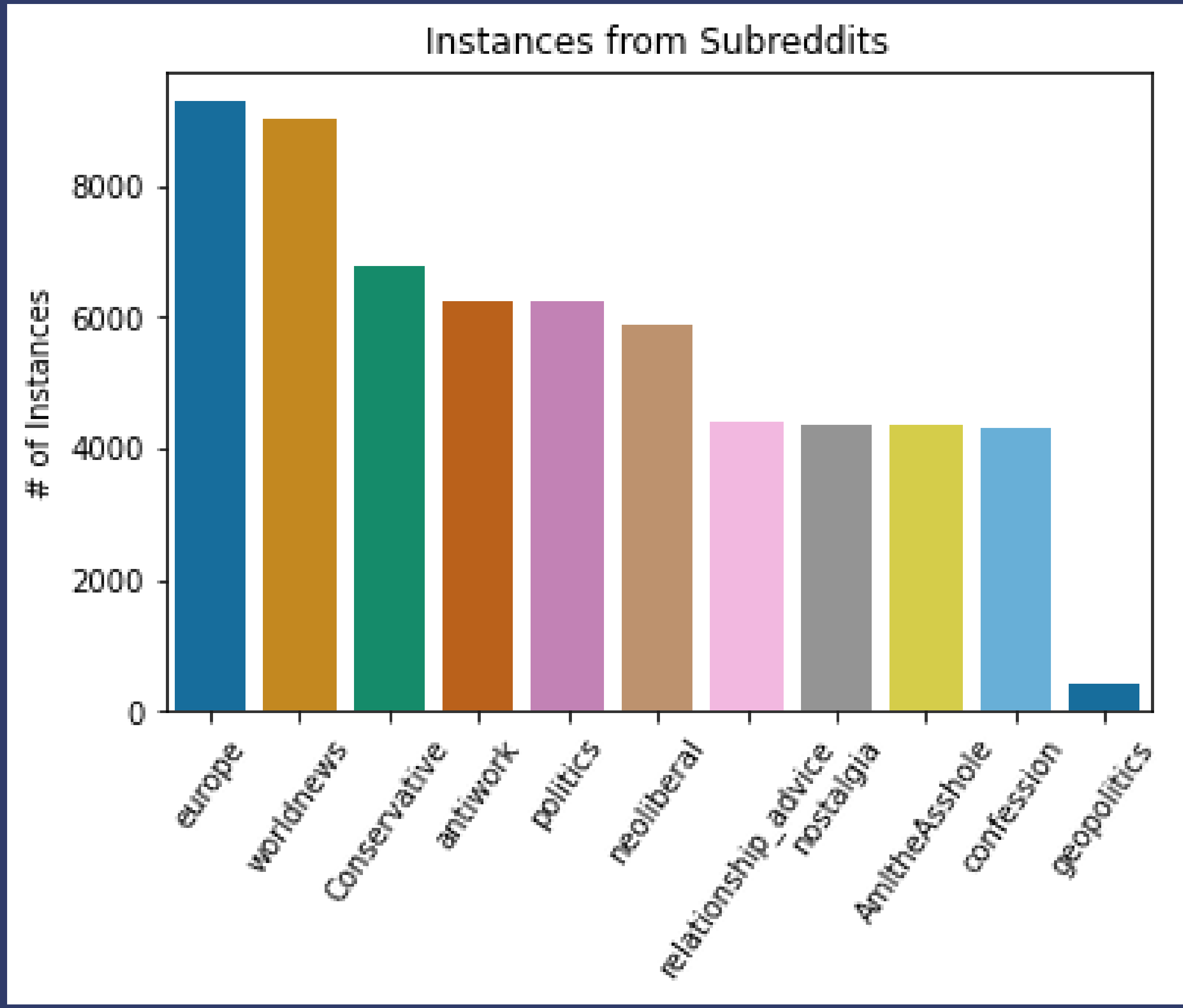
Sanctity/Degradation

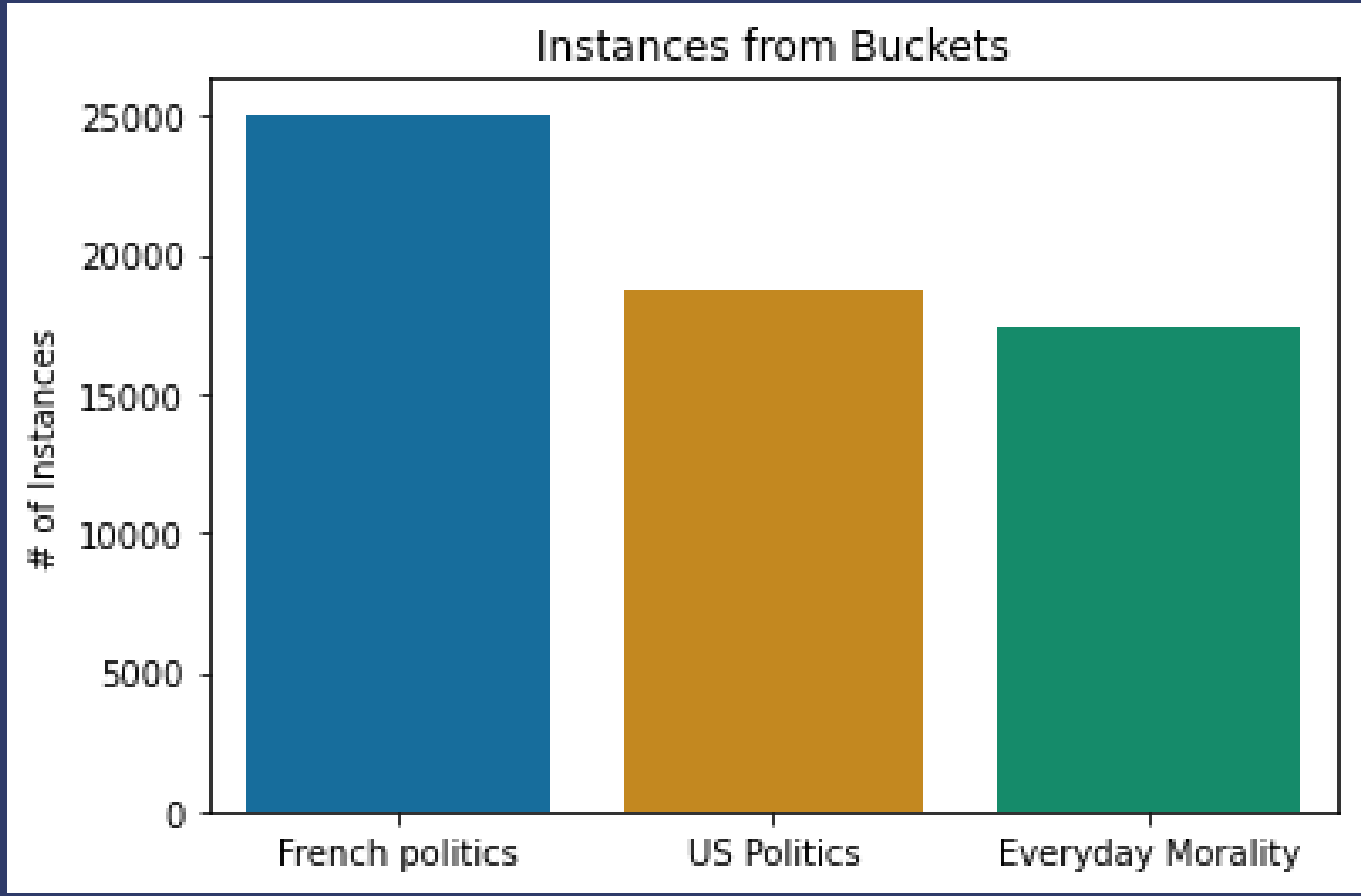
Datasets

Reddit Dataset

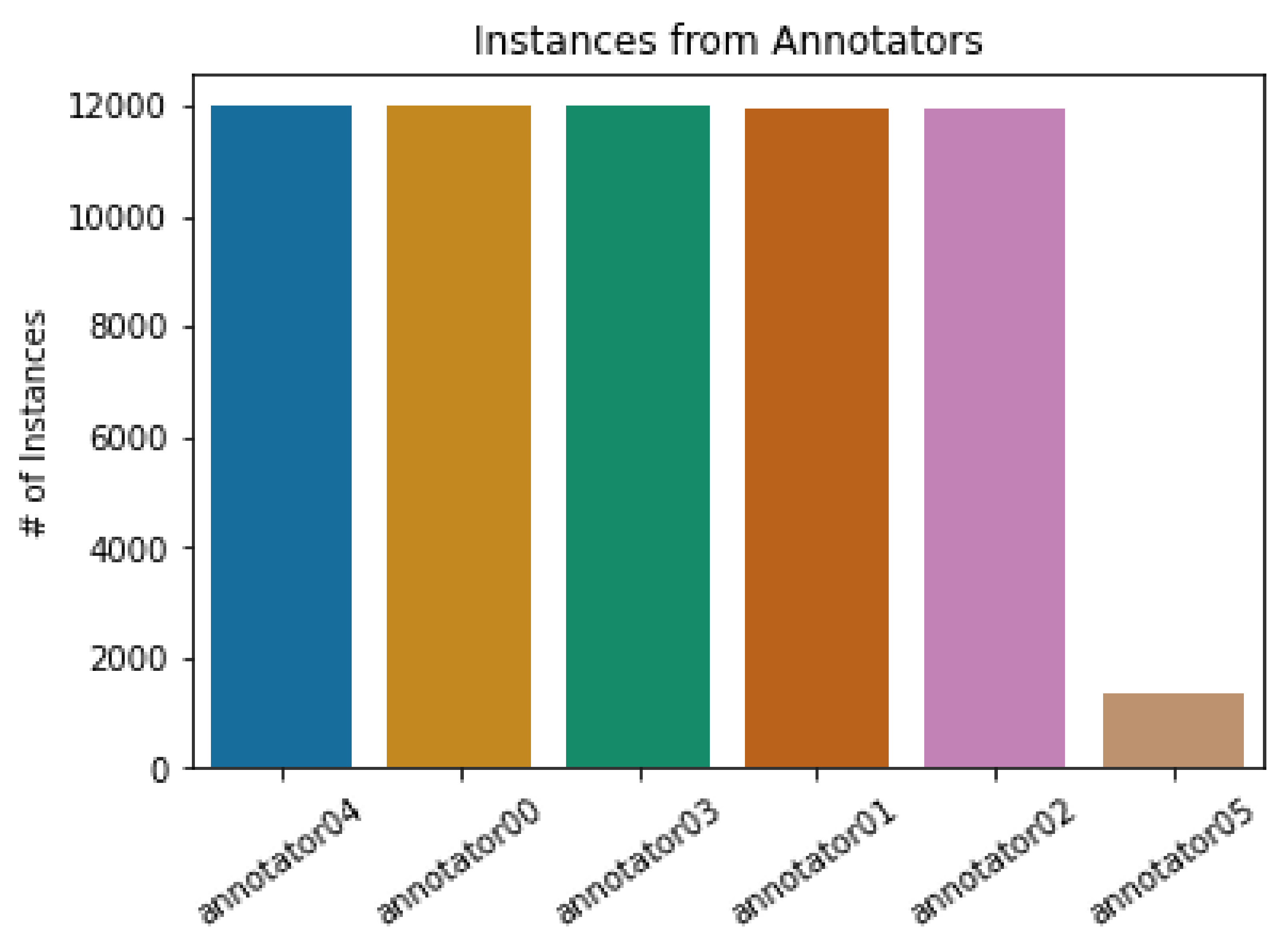
- 16,123 Reddit comments from 11 subreddits
- Hand-annotated for 8 classes
  - Care, Proportionality, Equality, Purity, Authority, Loyalty, Thin Morality, Implicit/Explicit Morality)

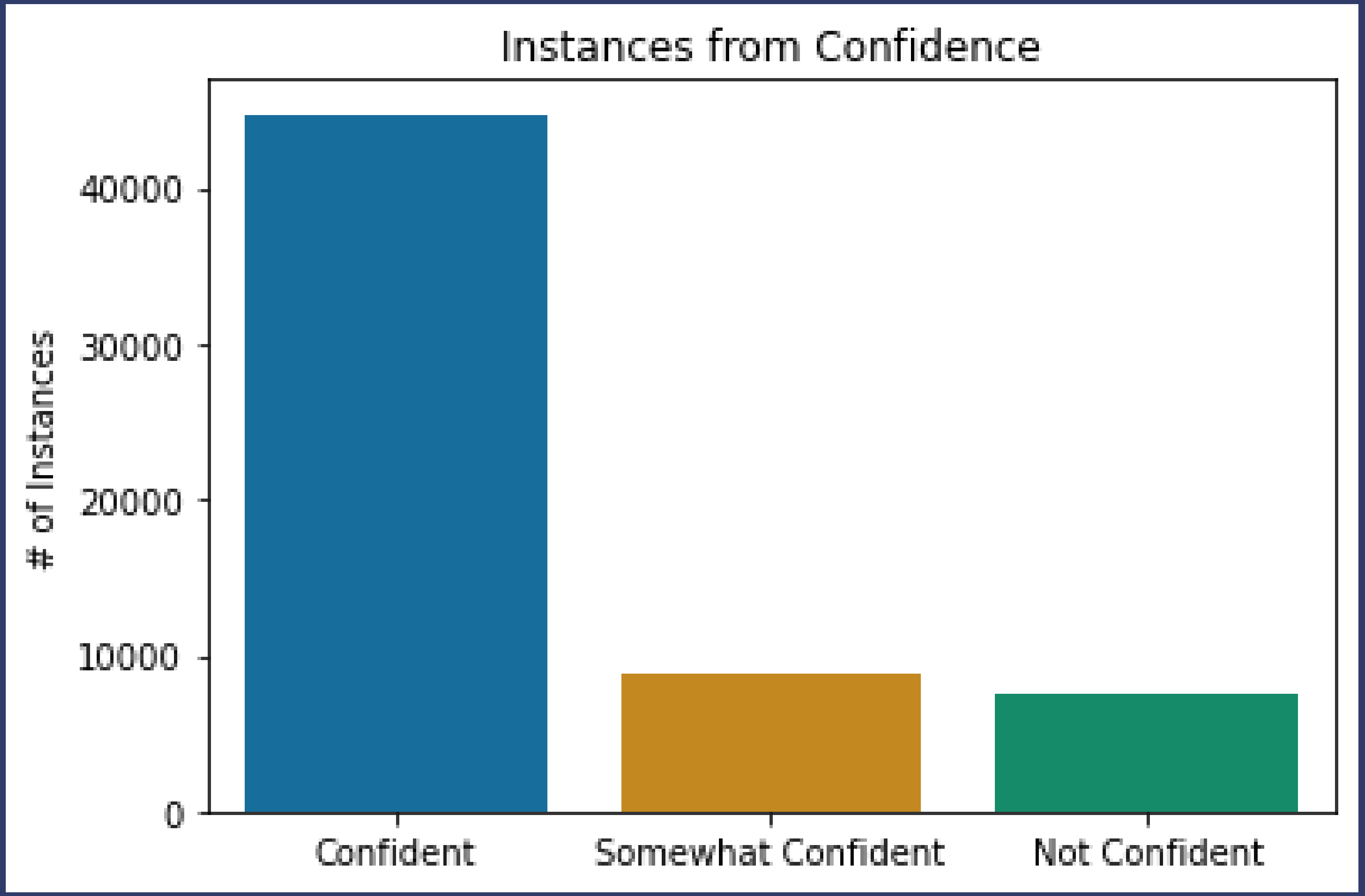
Datasets: Reddit Dataset



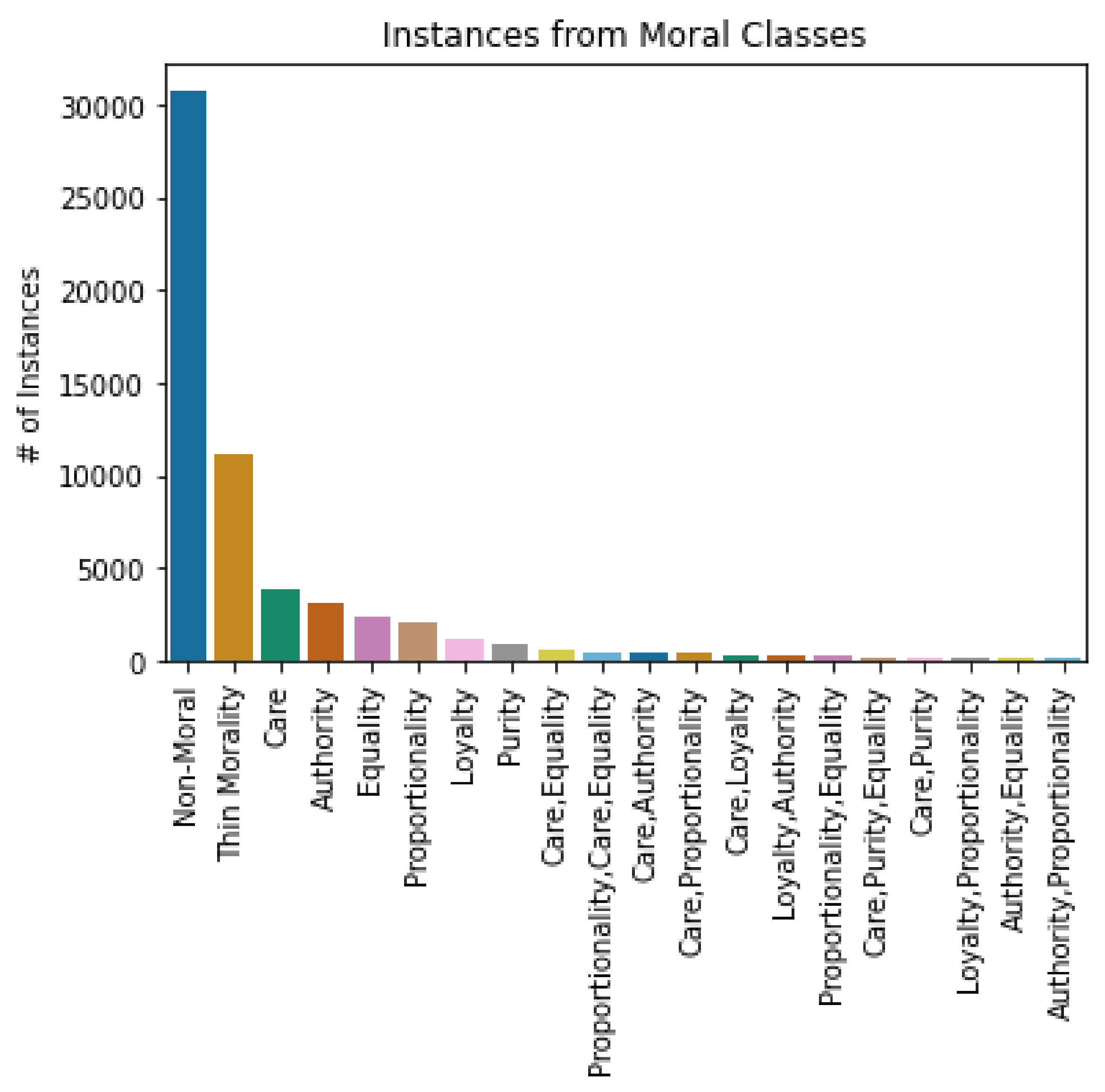


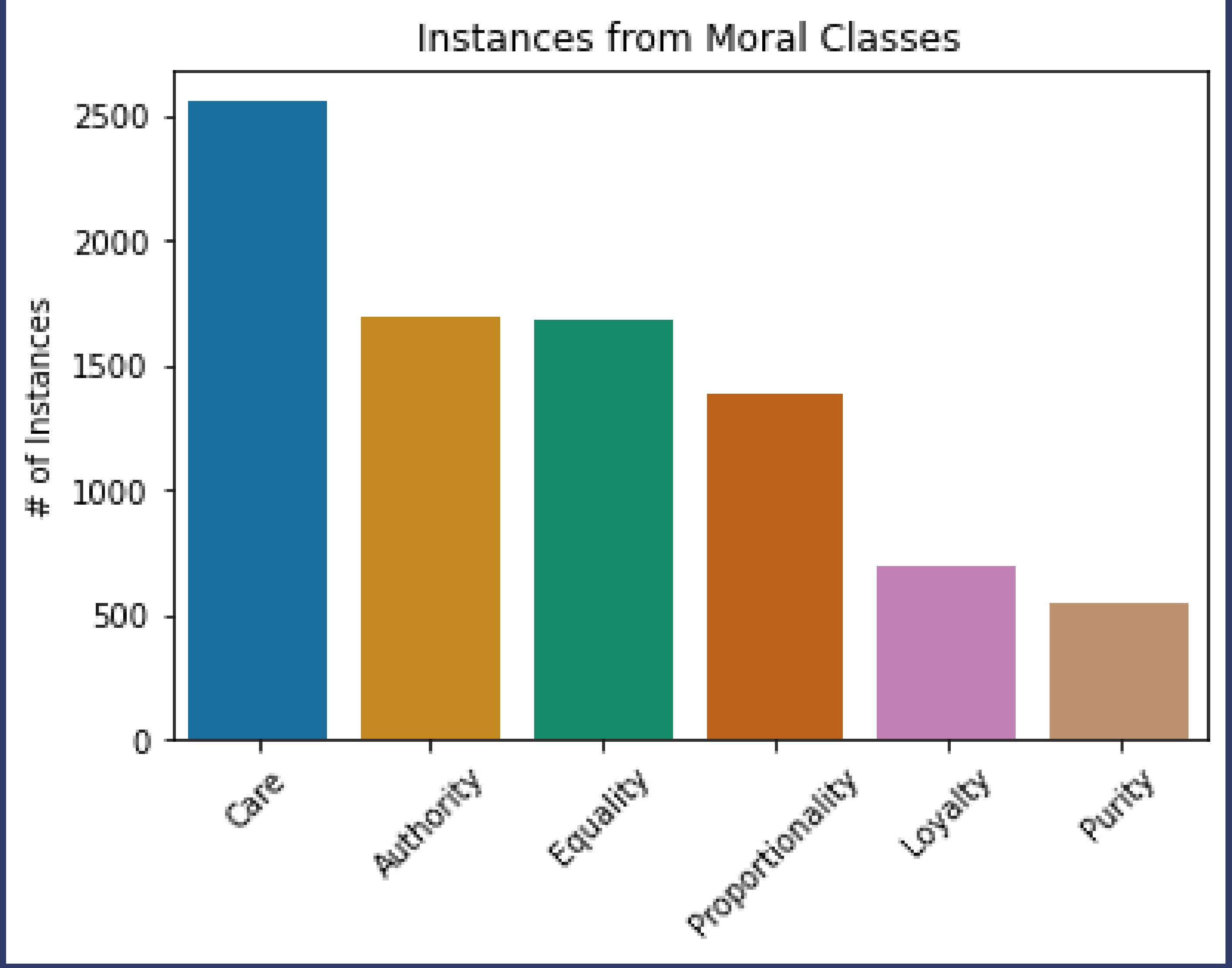
Datasets: Reddit Dataset





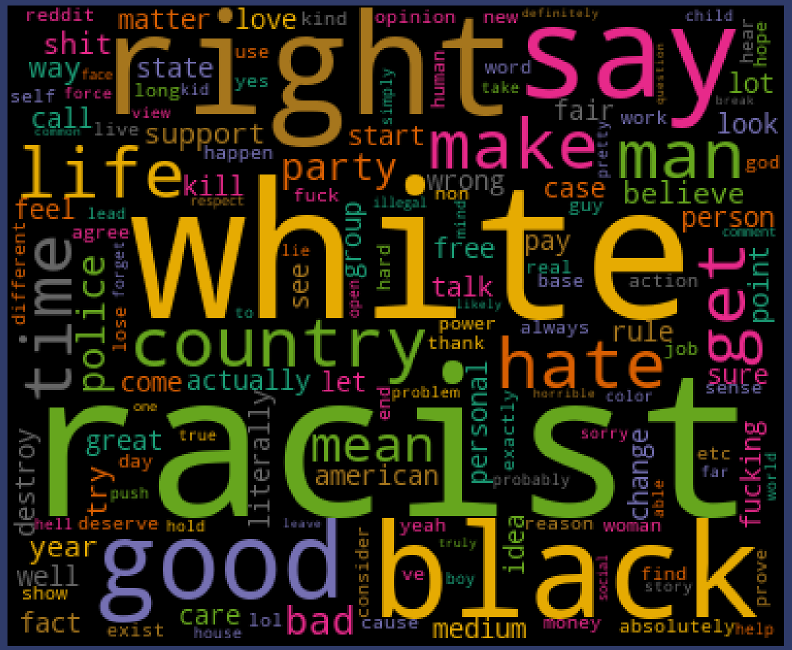
Datasets: Reddit Dataset





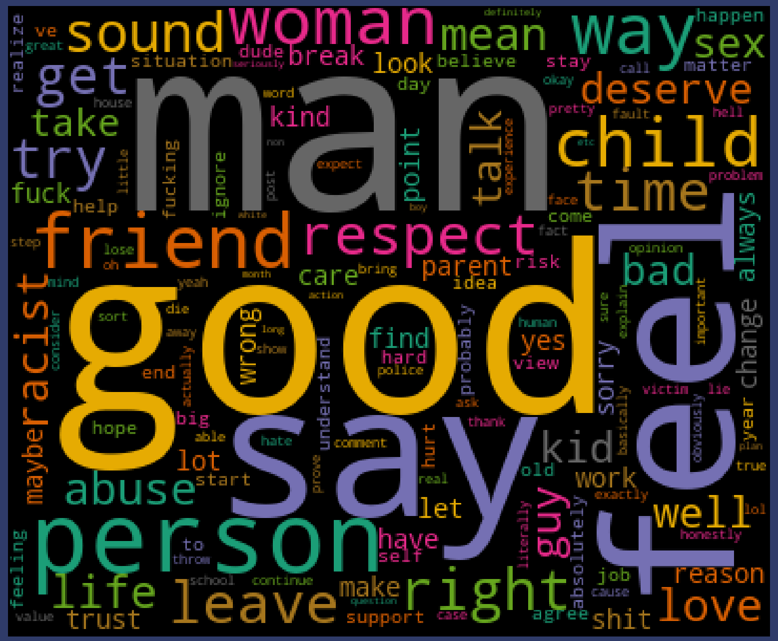
Datasets: Reddit Dataset

Conservative



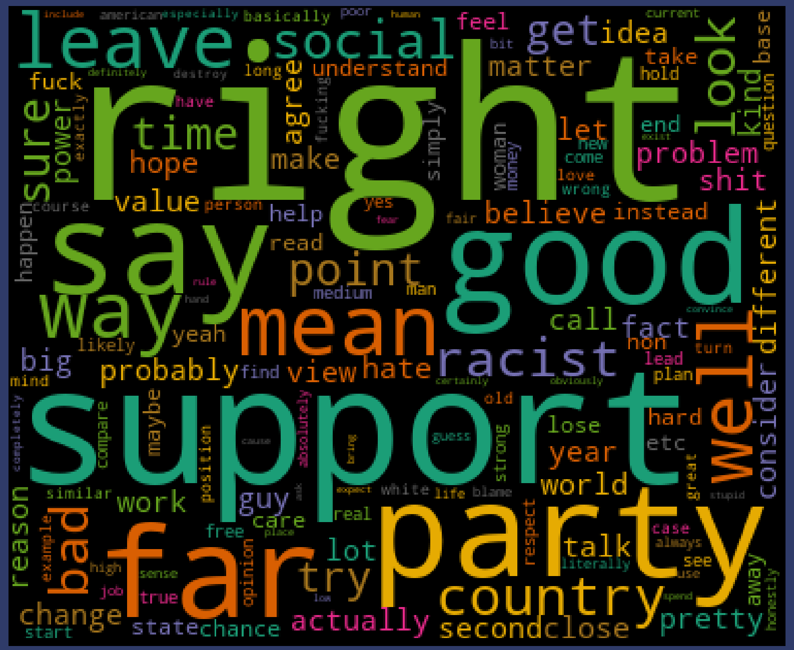
Am I The Asshole

Relationship Advice



Datasets: Reddit Dataset

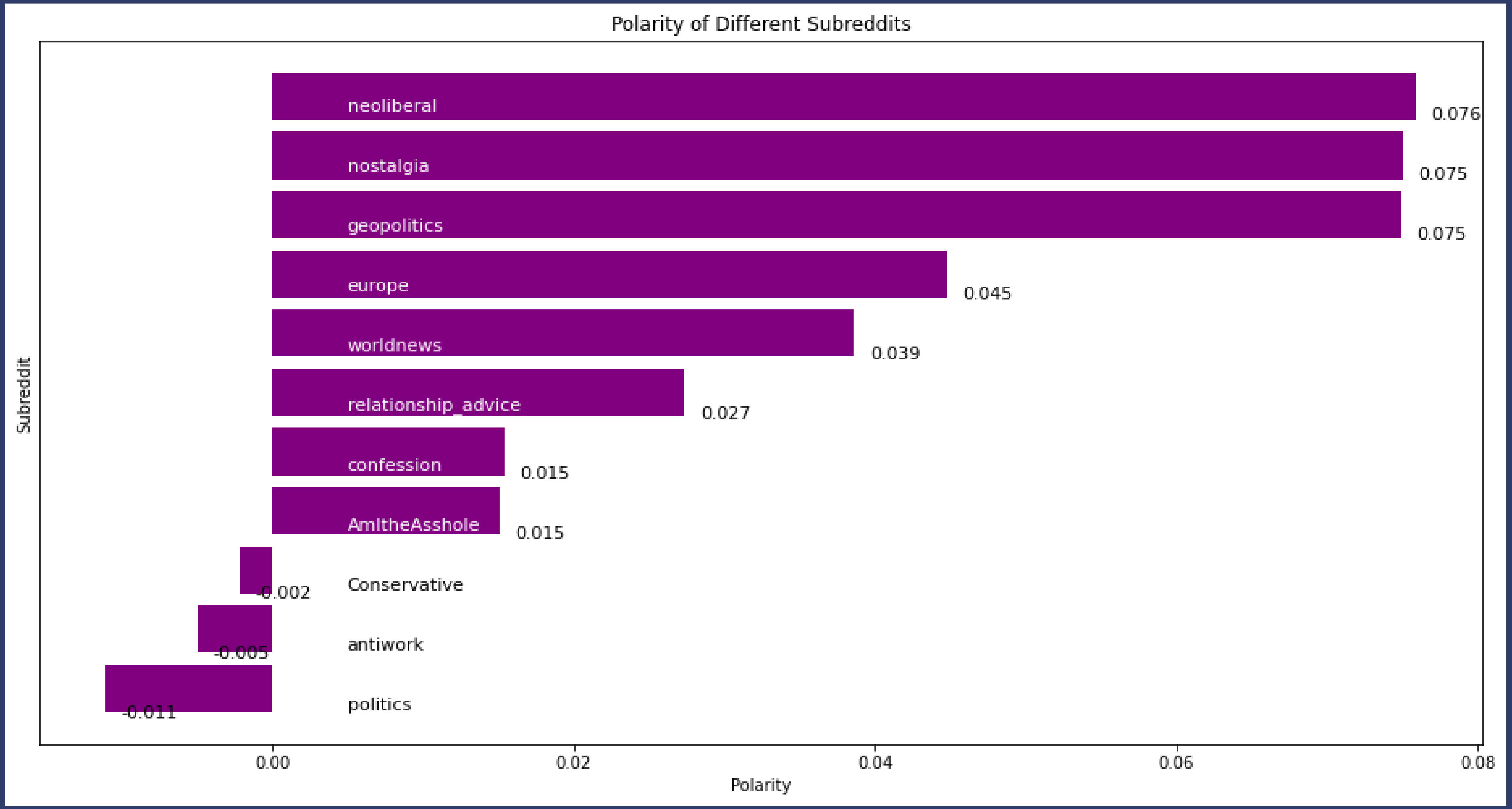
Europe



Neolibral

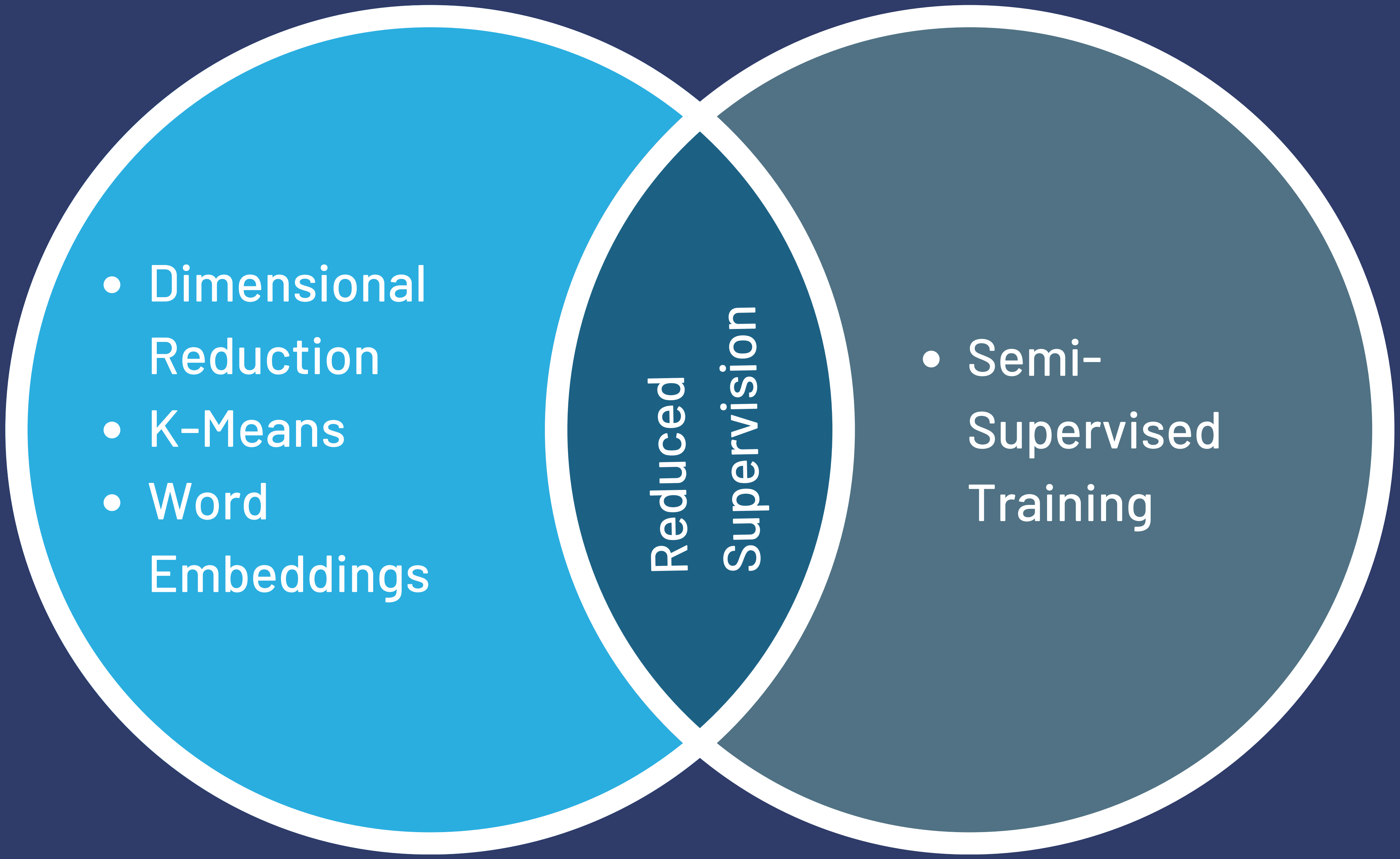
World News

Datasets: Reddit Dataset



Big Picture Roadmap

Can we uncover hidden morals without any labels?



Is training on a small amount of data and extrapolating to unlabeled data possible?

Unsupervised Baselines

Dimensional Reduction	KNN	Embedding Space

Embedding Space

- This model learns to represent data points in a higher-dimensional space.
- Using Word2Vec algorithm, the model learns word associates from a large group of texts. Once the dataset is trained, we can detect synonymous words or suggest additional words for a partial sentence.

	text	annotation	labeled_data	tokenized_vectors
0	MLP doesn't need to wait for a referendum to b...	Authority	3	[mlp, NOTneed, wait, referendum, break, europe...
1	Or - or - assclowns like Le Pen and Farage cou...	Equality	1	[assclowns, like, le, pen, farage, could, demo...
2	Congratulations on your victory Macron voters....	Care	0	[congratulations, victory, macron, voters, kno...

Embedding Space

Appearing together frequently in different sentences, Word2Vec understands that these words have some relationship.

```
# Training a Word2Vec model
keyed_vectors, keyed_vocab = w2v_trainer(df['tokenized_vectors'])
```

```
# Find the most similar words to "care/harm"
keyed_vectors.most_similar(positive=['care','harm'], negative=[], topn=15)
```

```
[('children', 0.9999149441719055),
 ('without', 0.9998925924301147),
 ('op', 0.9998899102210999),
 ('sorry', 0.9998860359191895),
 ('around', 0.9998852610588074),
 ('respect', 0.9998831748962402),
 ('nta', 0.9998764395713806),
 ('behavior', 0.9998753666877747),
 ('friend', 0.9998745322227478),
 ('live', 0.9998738765716553),
 ('man', 0.999873697757721),
 ('sex', 0.9998733401298523),
 ('parents', 0.999872088432312),
 ('others', 0.9998708963394165),
 ('wife', 0.9998690485954285)]
```

```
care_harm_concepts = ['care', 'benefit', 'amity','caring','compassion', 'empath', 'guard', 'peace', 'protect'
care_concepts = [concept for concept in care_harm_concepts if concept in keyed_vocab]
```

Embedding Space

```
def calculate_overall_similarity_score(keyed_vectors,
                                        target_tokens: List[str],
                                        doc_tokens: List[str]) -> float:

    target_tokens = [token for token in target_tokens if token in keyed_vectors]

    doc_tokens = [token for token in doc_tokens if token in keyed_vectors]

    if not (target_tokens and doc_tokens):
        return 0.0
    else:
        similarity_score = keyed_vectors.n_similarity(target_tokens, doc_tokens)
        return similarity_score
```

```
care_target_tokens:
fair_target_tokens:
loyal_target_tokens:
auth_target_tokens:
san_target_tokens: L
lib_target_tokens: L
doc_tokens: List[str
```

Embedding Space

lata	tokenized_vectors	tokenized_vectors_len	overall_care	overall_fair	overall_loyal	overall_auth	overall_san	overall_lib	overall_max_score	moral_foundations
3	[mlp, NOTneed, wait, referendum, break, europe...	53	0.997654	0.997627	0.997950	0.997999	0.997981	0.997319	0.997999	3.0
1	[assclowns, like, le, pen, farage, could, demo...	22	0.990250	0.990195	0.990874	0.990962	0.990964	0.989595	0.990964	4.0
0	[congratulations, victory, macron, voters, kno...	36	0.997246	0.997210	0.997566	0.997612	0.997598	0.996890	0.997612	3.0
1	[german, constitution, NOTlet, hitler, become,...	13	0.976155	0.976087	0.977143	0.977279	0.977298	0.975132	0.977298	4.0

Embedding Space

```
# OSSA Model Evaluation
print("OSSA Model Evaluation: ")
evaluate_model(df['labeled_data'],
               df['moral_foundations'])

print("========================")
```

```
OSSA Model Evaluation:
* Accuracy Score:  21.8966%
* F1 Score:  21.8966%
* Recall Score:  21.8966%
* Precision Score:  21.8966%
========================
```

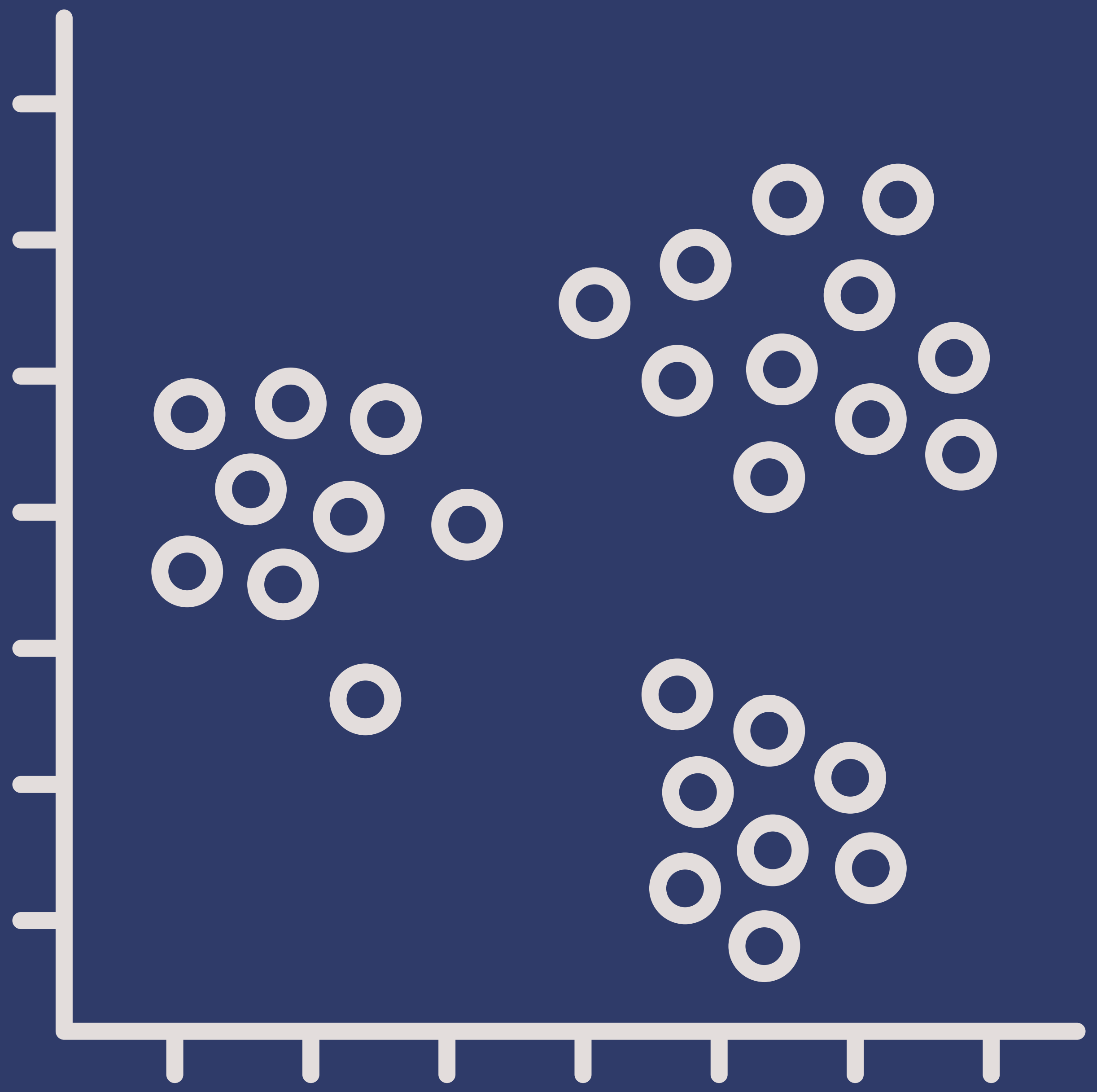
KMeans Clustering

Clustered on

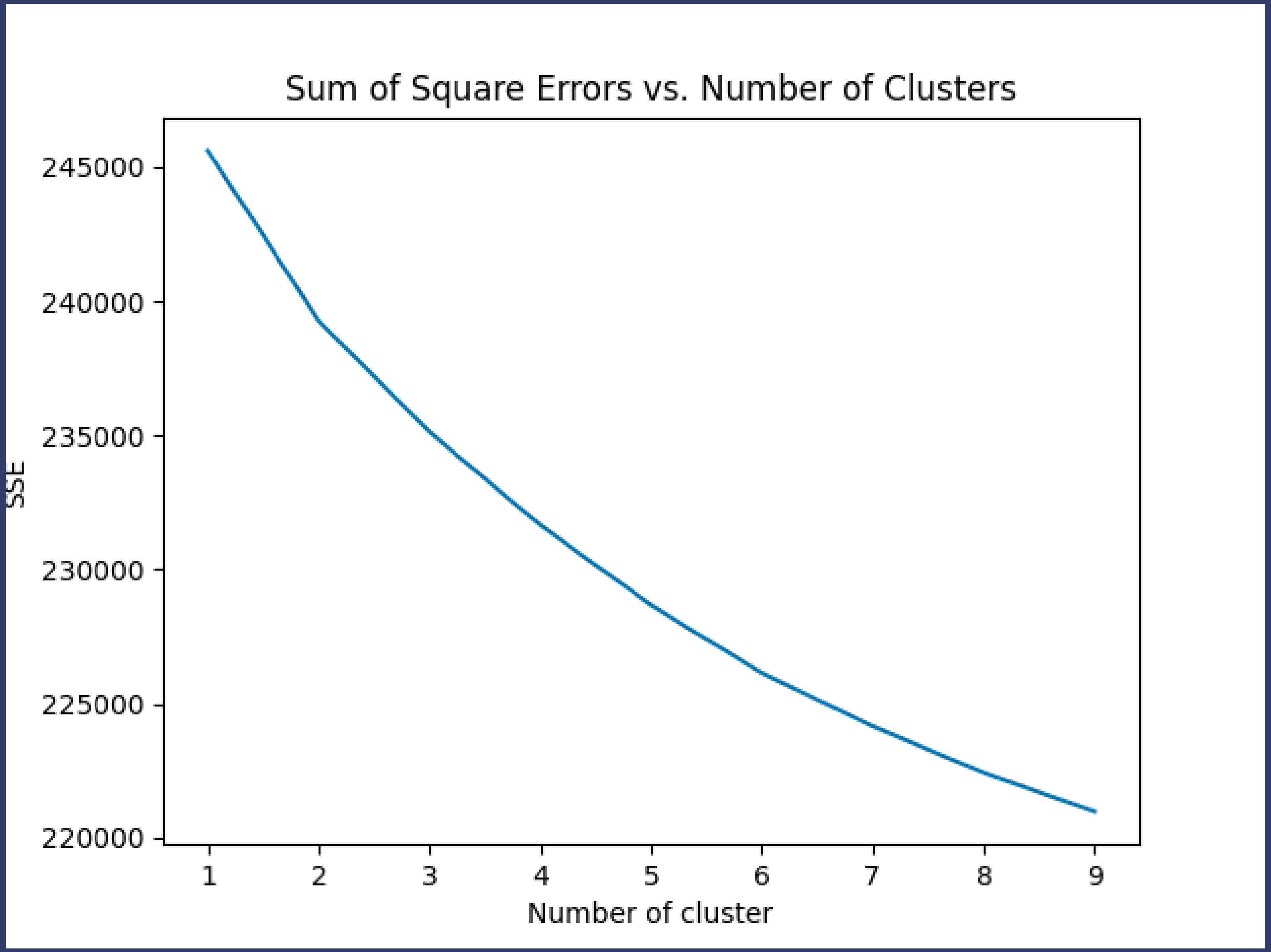
- Word2Vec Embedding
- Top n PCA components of Word2Vec Embedding
- TF-IDF Embedding

Evaluation Metrics

- Davies-Bouldin Index
- Callinski-Harabasz Index



KMeans Clustering: Word2Vec Embedding



Davies Bouldin Score: 3.94
Calinski Harabasz Score: 129.20

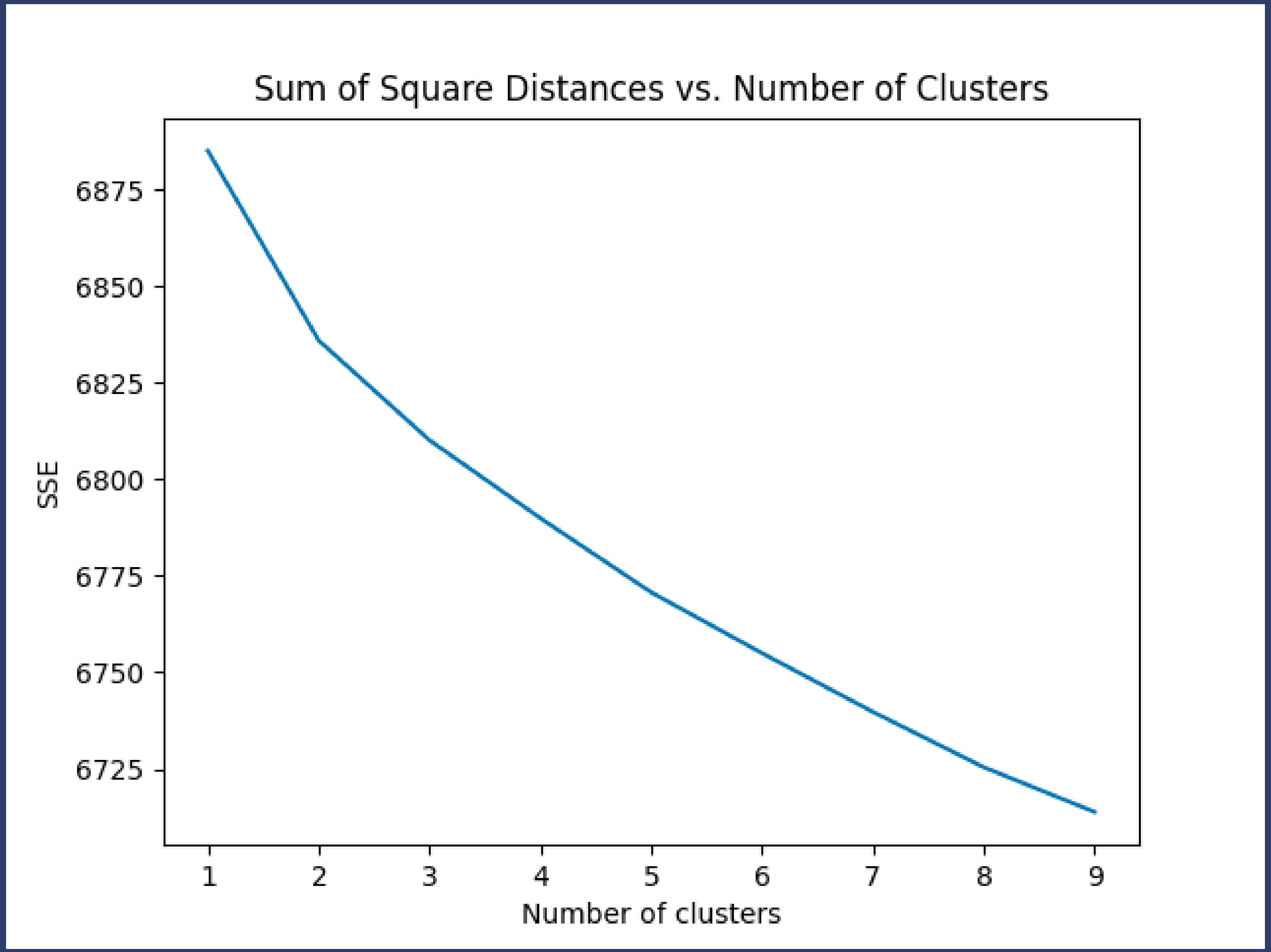
KMeans Clustering: Word2Vec Embedding

```
 ## Words found in cluster 0 ##
[('nationalistic', 0.7419905662536621),
 ('nationalism', 0.7199922800064087),
 ('encompass', 0.7065395712852478),
 ('regime', 0.6837404370307922),
 ('patriotic', 0.682215690612793),
 ('authoritarian', 0.6742655634880066),
 ('totalitarian', 0.6666761636734009),
 ('intolerance', 0.661752164363861),
 ('theocratic', 0.6608919501304626),
 ('uphold', 0.6589691638946533),
 ('secularism', 0.6566293239593506),
 ('xenophobic', 0.6544418334960938),
 ('colonial', 0.652903139591217),
 ('marxist', 0.6501309275627136),
 ('ideological', 0.648766040802002),
 ('embrace', 0.6481760740280151),
 ('isolationist', 0.6434253454208374),
```

```
 ## Words found in cluster 1 ##
[('communicate', 0.6528956890106201),
 ('immature', 0.6414459347724915),
 ('interact', 0.6256312727928162),
 ('selfish', 0.6238377094268799),
 ('inappropriate', 0.6223106384277344),
 ('behaviour', 0.6155641078948975),
 ('emotionally', 0.6113891005516052),
 ('hormonal', 0.6111546754837036),
 ('pester', 0.6094121932983398),
 ('judgment', 0.6081652641296387),
 ('obtuse', 0.6076530814170837),
 ('stmake', 0.6041022539138794),
 ('manipulative', 0.6001163721084595),
 ('apologize', 0.5996392965316772),
 ('sexually', 0.5986785292625427),
 ('disrespect', 0.5914411544799805),
 ('autonomy', 0.591350257396698),
 ('deviant', 0.5903379917144775),
 ('disrespectful', 0.5876566767692566),
```

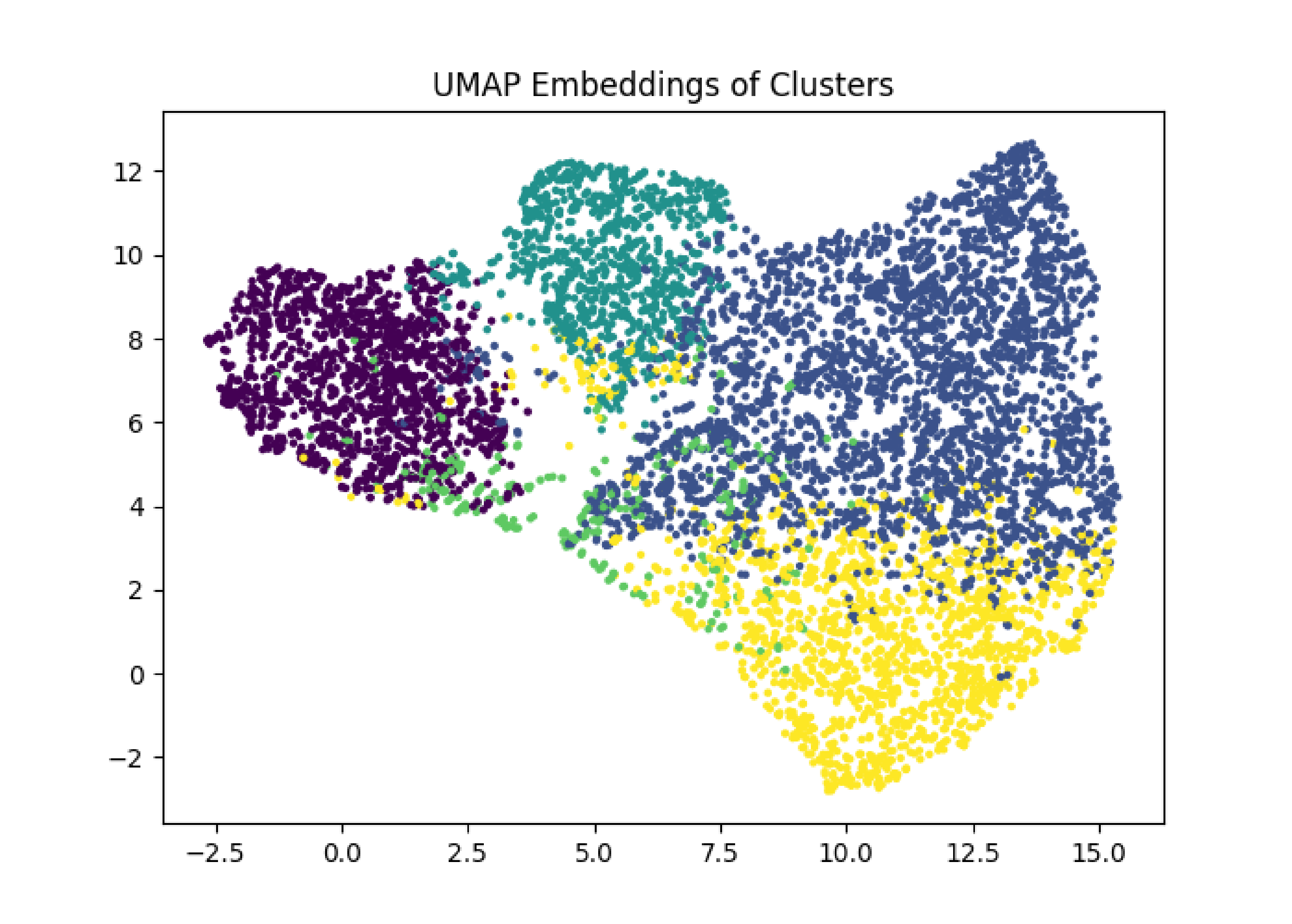
```
 ## Words found in cluster 2 ##
[('payment', 0.7700842022895813),
 ('loosen', 0.75533527135849),
 ('vulnerability', 0.7356607913970947),
 ('revenue', 0.7337859272956848),
 ('contract', 0.7334370613098145),
 ('disposable', 0.7294864058494568),
 ('scarcity', 0.726427376270294),
 ('surplus', 0.7195096611976624),
 ('ip', 0.7135844230651855),
 ('prisoner', 0.7130150198936462),
 ('innovation', 0.7127234935760498),
 ('allocate', 0.711272060871124),
 ('ppe', 0.7076494693756104),
 ('productivity', 0.7066180109977722),
 ('warfare', 0.7039303779602051),
 ('macroeconomic', 0.6943965554237366),
 ('regulate', 0.6937347054481506),
 ('ownership', 0.6932600140571594),
 ('spending', 0.6919416785240173),
 ('training', 0.6910233497619629),
 ('maximum', 0.69019562005967),
 ('supply', 0.6887563467025757),
```

KMeans Clustering: Topic Modelling (LSA)



Davies Bouldin Score: 10.78
Calinski Harabasz Score: 29.62

KMeans Clustering: Topic Modelling (LSA)



people, vote, right, trump, racist, france, eu

France, fascist, party, right, trump, nazi, marine
Anti, nationalist, election, putin, nationalism,
immigration

French, leader, president, candidate,
election, liberal

Unsupervised Baselines

K-Means Clustering

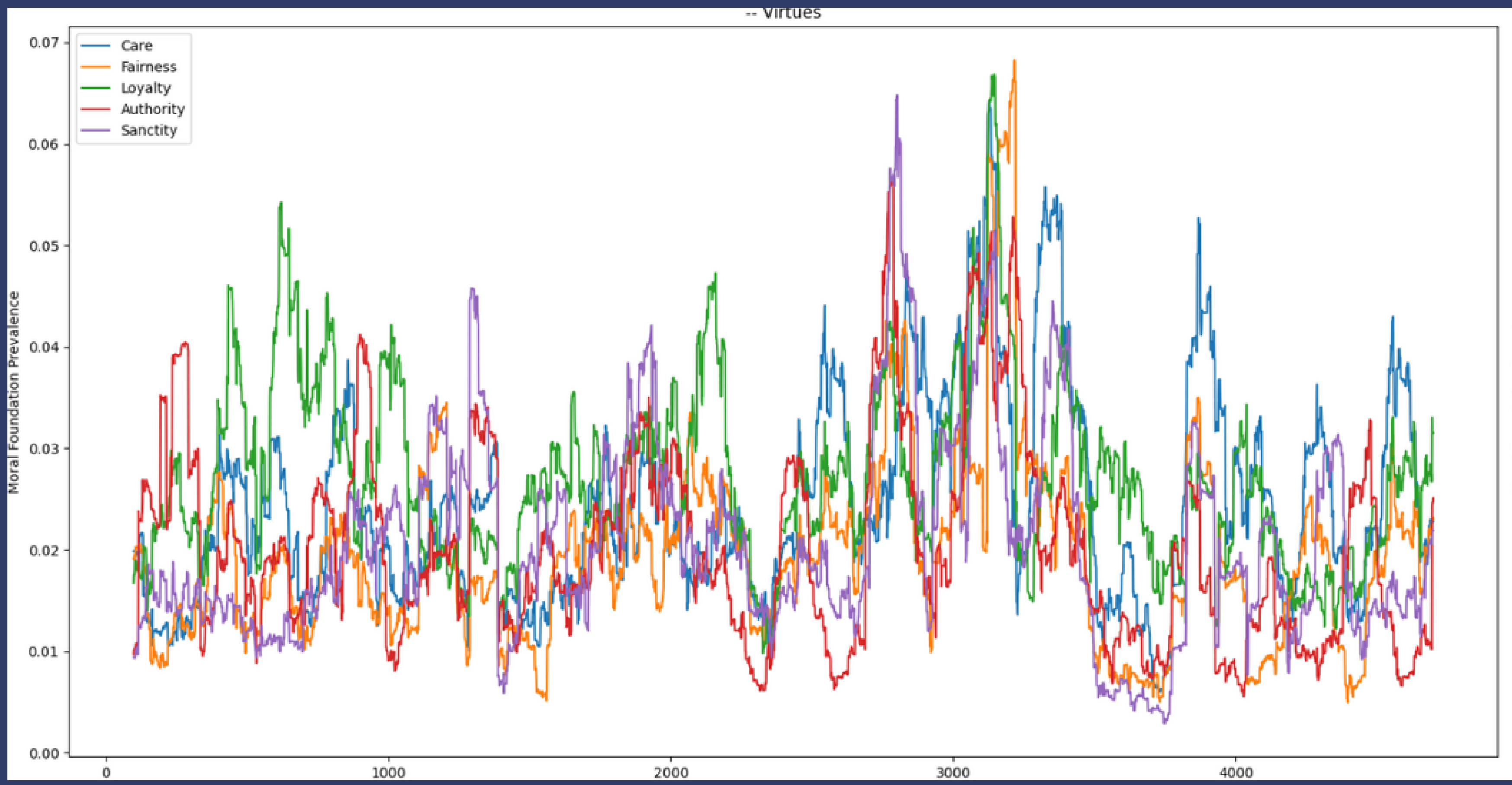
- Difficult to connect word embeddings back to sentences
- Unable to map reddit posts directly to morality, confined to work analysis
- Clusters are not entirely separable

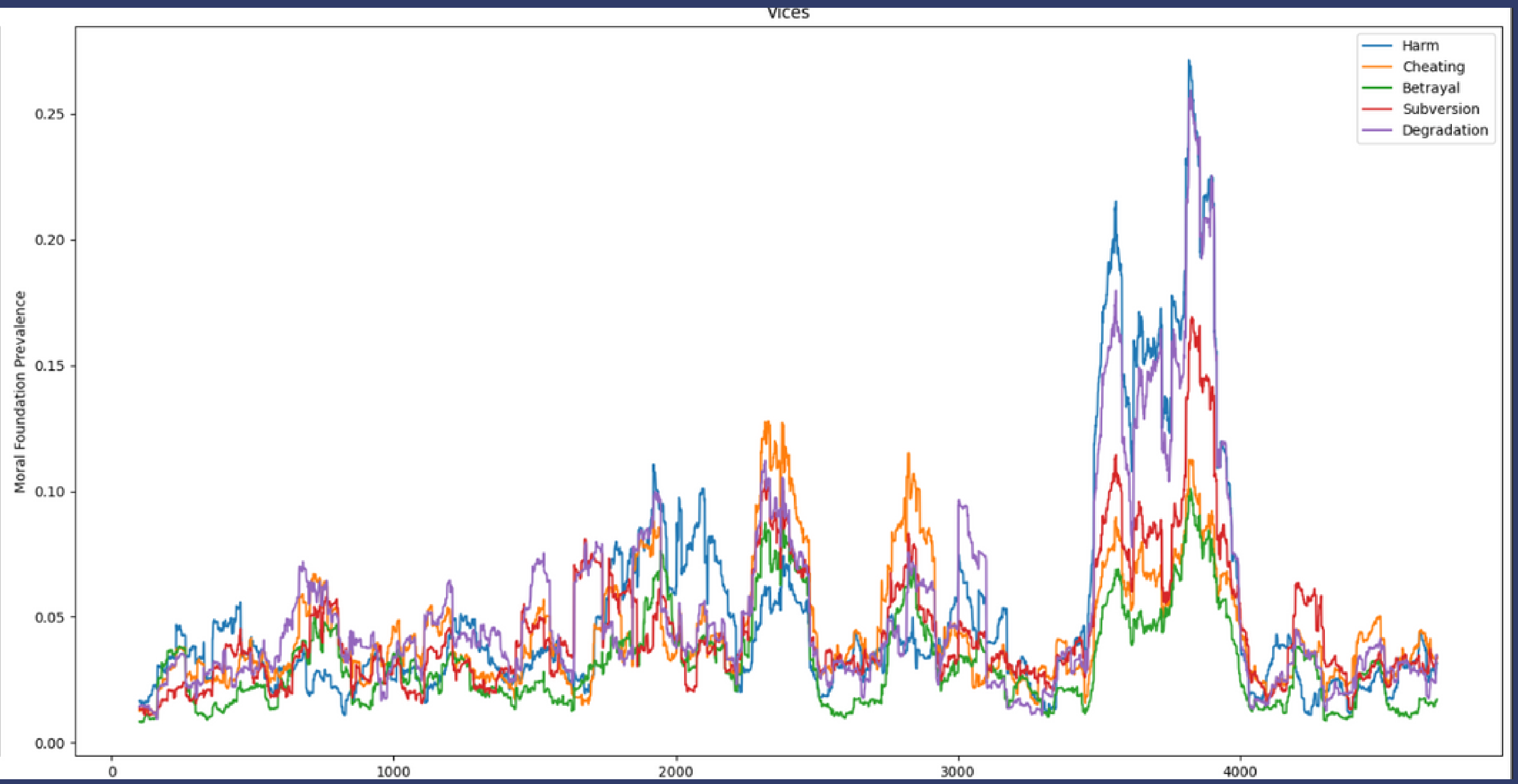
Embedding Space

- Worked on classifying at least one of the moral foundations
- Did not work was comparing the text into the right moral concept,

Semi-Supervised

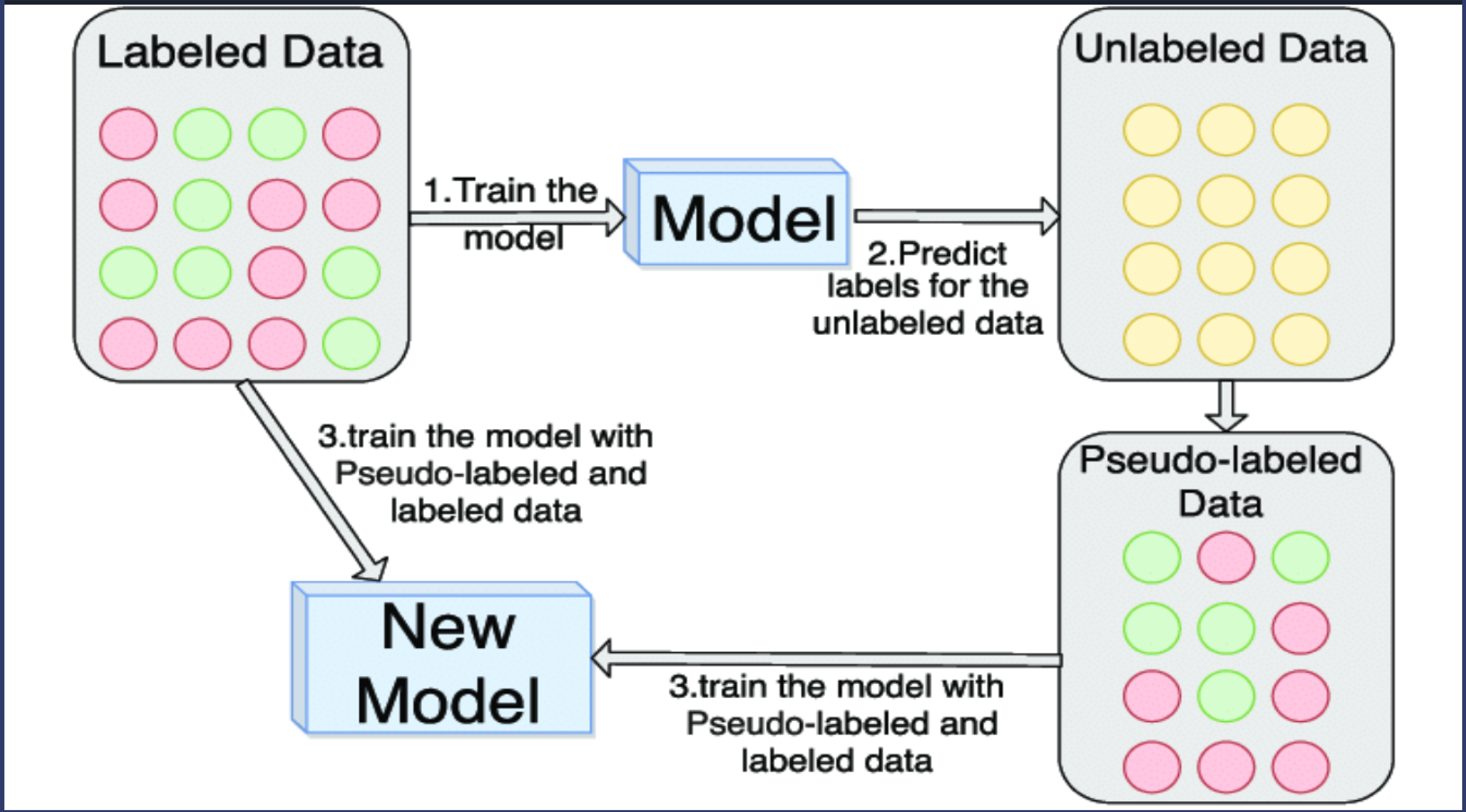
Sentiment Distribution





- "I have an adult brother with Down Syndrome who lives independently in an apartment with staff who assist him. Intellectual disability ranges vastly in adults with Downs and it's impossible to know whether or not he knew the value of the money he handed you. I would 100% double-check to make sure that he realized how much he handed you. Individuals with Downs are often very kind and it's possible he knew it would make you happy but didn't realize the repercussions of his kindness.

- "It's not illegal. You can't get a mortgage though. Also you might get investigated for money laundering."


Labeled Data
1.Train the model
Model
2.Predict labels for the unlabeled data
Unlabeled Data
Pseudo-labeled Data
3.train the model with Pseudo-labeled and labeled data
New Model
3.train the model with Pseudo-labeled and labeled data

Conclusion

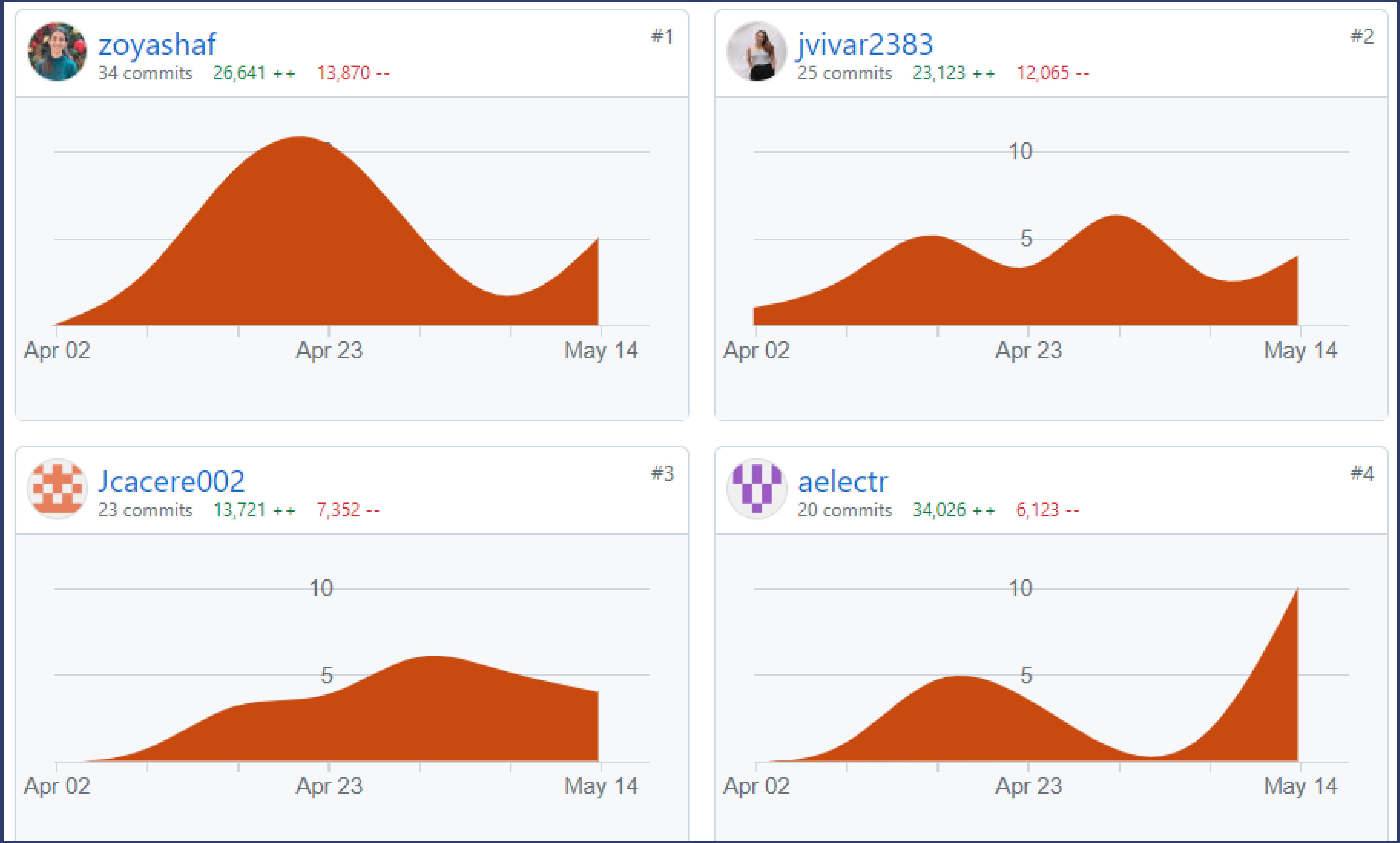
Model Performance

- Unsupervised models were hard to extrapolate back to morals and achieved low accuracy
- A semi-supervised method performed well even with a small amount of data

General Remarks

- Working with unlabelled data requires a lot more engineering and feature selection than working with labeled data
- Quality of data is really important

Thank you!


zoyashaf
34 commits 26,641 ++ 13,870 --
#1
Apr 02
Apr 23
May 14
jvivar2383
25 commits 23,123 ++ 12,065 --
#2
10
5
Apr 02
Apr 23
May 14
Jcacere002
23 commits 13,721 ++ 7,352 --
#3
10
5
Apr 02
Apr 23
May 14
aelectr
20 commits 34,026 ++ 6,123 --
#4
10
5
Apr 02
Apr 23
May 14