Classifying Propositional Content

In Argumentative Discourse Units

M.EIC - Natural Language Processing
April, 21. 2022

Matheus Santos (up202111214@edu.fe.up.pt)
Telmo Baptista (up201806554@fe.up.pt)

Table of contents

01

Problem Overview

What is the problem and
what we know about it

Classifying ADUs

Our objective is to classify the propositional content of ADUs from the Público newspaper into:

- Facts;
- Policies;
- Negative Value;
- Neutral Value;
- Positive Value;

Articles

Contains keywords

Topic of article

Other information

ADUs

Contains the ADU excerpts

02

Data Analysis

Exploring the data available

ADU Analysis

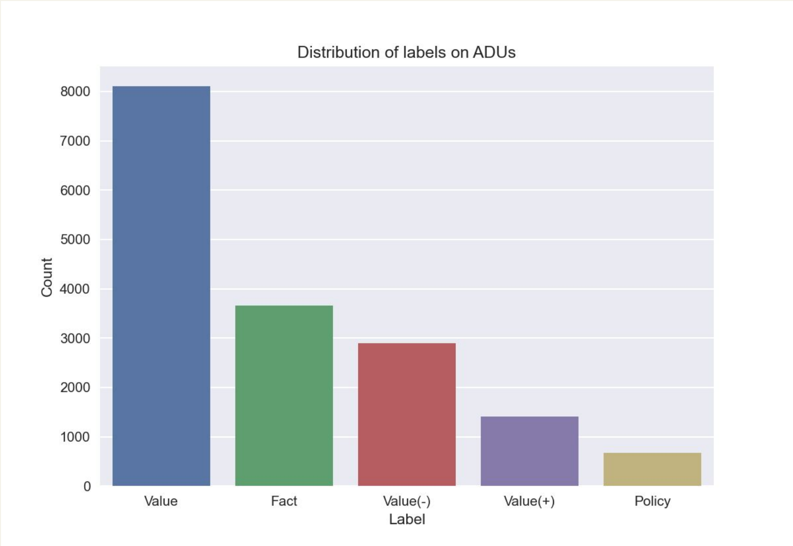


Fig. 1 - Label distribution on the dataset

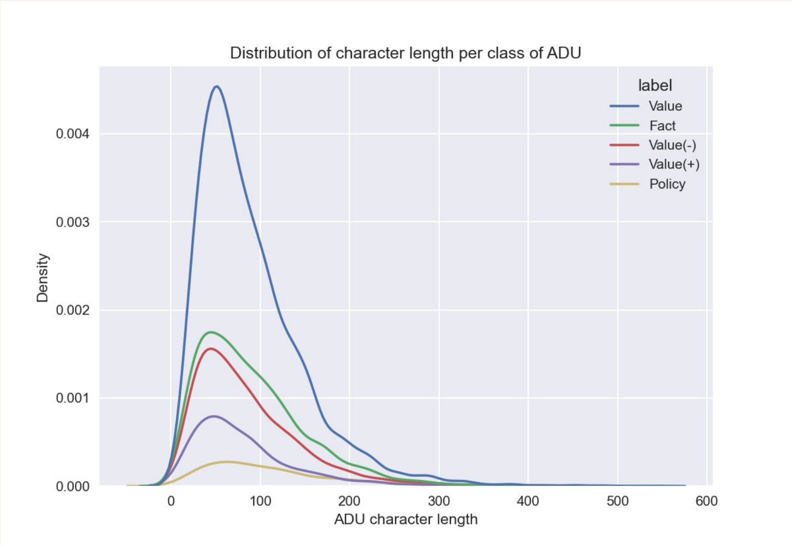


Fig. 2 - Distribution of character length of ADUs on the dataset

Frequent Words

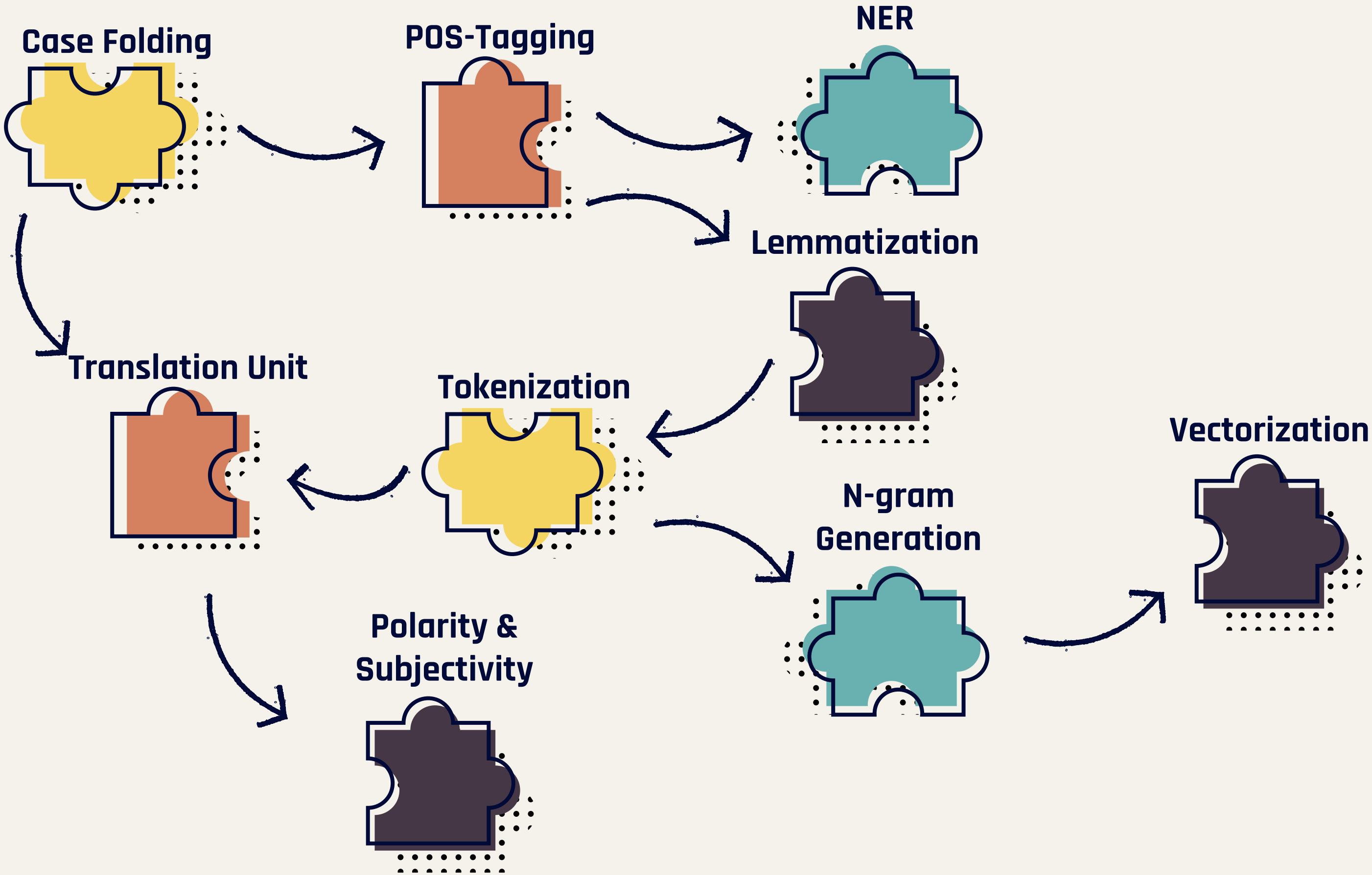
Fact	Policy	Value (-)	Value	Value (+)
direito, ano	**dever**, necessário, precisar, deixar, medir, estratégia	não, **mau**, **problema**	(shares lots of words with all, and none are extremely noticeable)	**importante**, **melhor**, permitir, melhorar, bom, qualidade

03

Processing

Transforming the data

Token Processing Unit



Polarity of a sentence

Não foi mau → 0.23 | 0.43 | 0.34 → 0.24

But how?

Eu não odiei, mas não foi impressionante.

	não		odiei		mas		não		impressionante	
	1.0		1.0		0.0		1.0		0.1	
	0.0	→	0.0	→	1.0	→	0.0	→	0.3	> -0.389
		-1.0		-0.25		-0.125		-0.5625		
	0.0		0.0		0.0		0.0		0.6	
	1.0	→	-0.5		-0.5		1.0		-0.5	

What happened?

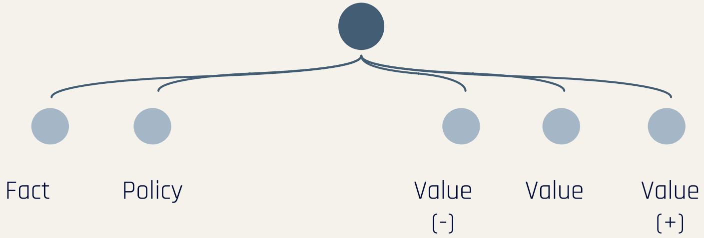
04

Classification Task

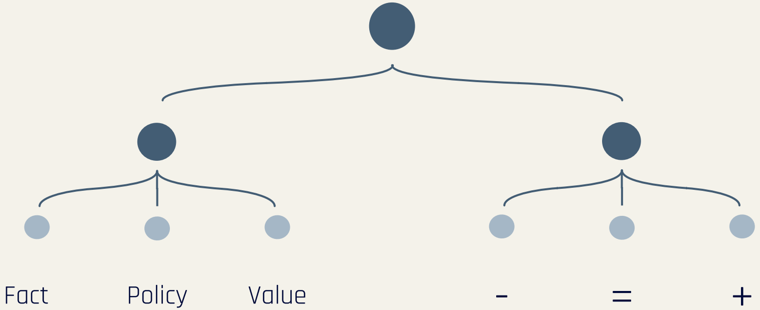
What tasks were
considered

Two Approaches

Multi-class



Multi-output multi-class



Exploring Features

Fact

milhao euro
nao haver
ensinar superior
nao ser
direito humano
medio europeu
ter ser
ultimo ano
assembleia republica

Policy

nao poder
nao esquecer
sociedade civil
publico privar
nao deixar
emprestimo fraudulento
nao ficar
poder continuar
futebol portugues

Value (-)

nao haver
nao ser
nao deixar
portugal nao
nao existir
servico publico
empregar cientifico
programar empregar
nao resolver

Value

nao haver
nao ser
nao ter
uniao europeu
ensinar superior
estar unir
ter ser
nao existir
ultimo ano

Value (+)

direito humano
nao deixar
ultimo ano
bom noticia
valer peno
barragem alvito
protejo tejo
resolver problema
papel importante

	token_len	n_entities	unique_entities	org_entities	loc_entities	per_entities	misc_entities	adu_polarity	token_polarity	blob_polarity	blob_subjectivity	adj_count	adv_count	cconj_count	sconj_count	noun_count	det_count	verb_count	intj_count	part_count	pron_count	propn_count	punct_count	polarity	label	value_type
label	-0.05	-0.08	-0.07	-0.03	-0.08	-0.02	-0.05	0.00	-0.01	0.01	0.09	-0.00	0.04	0.00	0.00	-0.05	0.00	0.02	-0.00	-0.01	0.02	-0.08	-0.02	0.00		
value_type	-0.01	0.02	0.02	0.00	0.02	-0.00	0.02	0.30	0.23	0.21	0.02	0.00	-0.04	-0.02	0.01	-0.00	-0.00	-0.02	0.00	0.00	-0.01	0.02	0.00	0.33	-0.10	

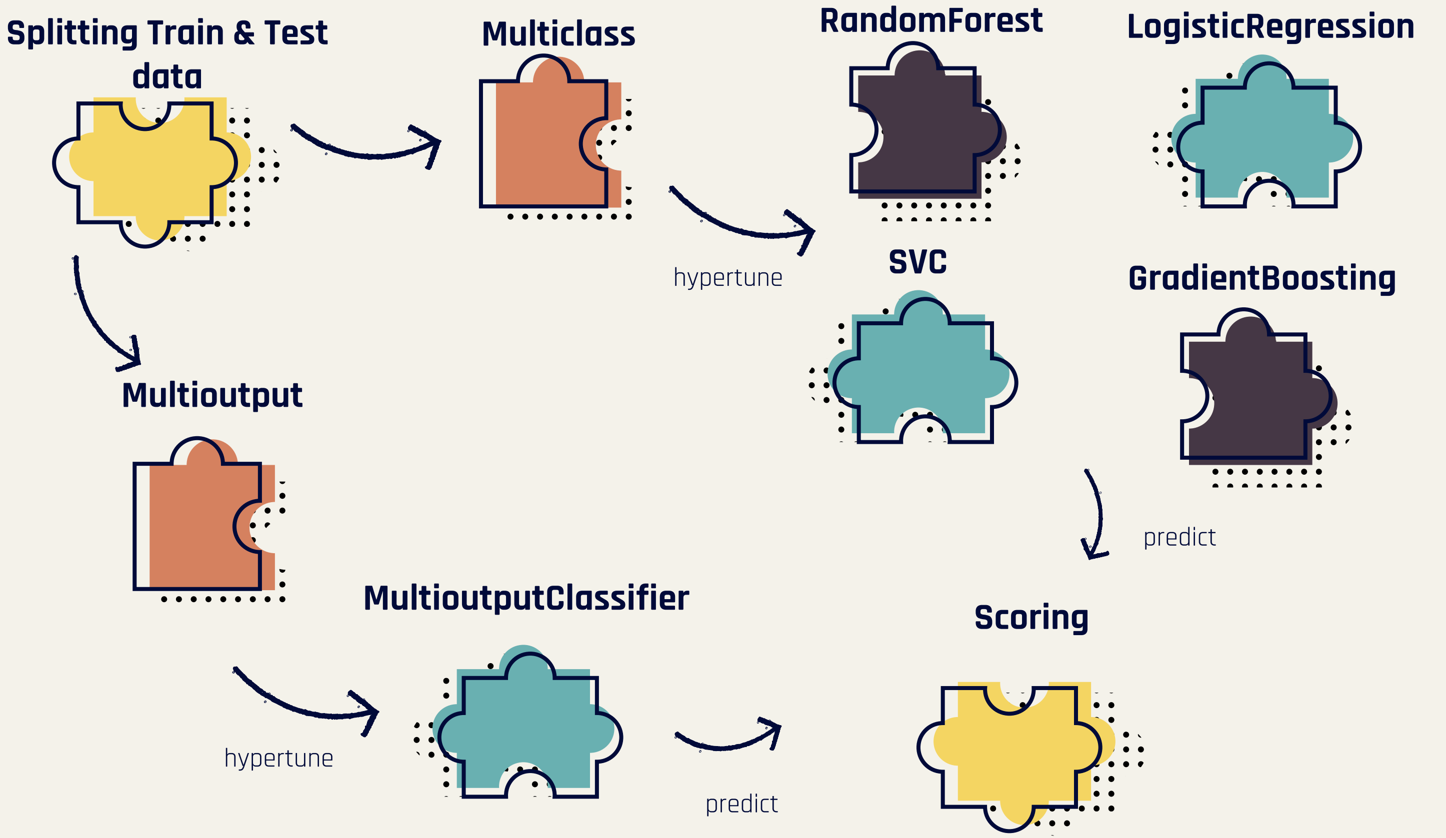
Fig. 3 - Correlation of extra features with our target columns (label: Fact, Policy, Value; value_type: Negative, Neutral, Positive)

05

Modelling

What techniques and
algorithms were applied?
How it was trained?

Modelling Process



Hypertuning Process

Cross-Validation

Models are trained based on 5-fold cross validation

Metric

Models are scored based on *f1_score*

GridSearch

Hypertuning based on grid search over a list of parameters

06

Results

What were the results?
And what information do
we get from it?

Results of (1+2)grams

	Base	Random Forest	Logistic Regression	Gradient Boosting	SVC	MOMC
Precision	0.32	0.46	0.54	0.50	0.53	0.51
Recall	0.19	0.40	0.52	0.51	0.54	0.51
F1-Score	0.22	0.41	0.44	0.46	0.51	0.45

Table. 1 - Weighted averaged results for various models on input with unigrams and bigrams

	Fact	Policy	Value (-)	Value	Value (+)
Precision	0.47	0.67	0.55	0.55	0.51
Recall	0.30	0.40	0.38	0.76	0.30
F1-Score	0.37	0.50	0.45	0.63	0.38

Table. 2 - Results per class for SVC model on input with unigrams and bigrams

Model Performance

SVC Performance

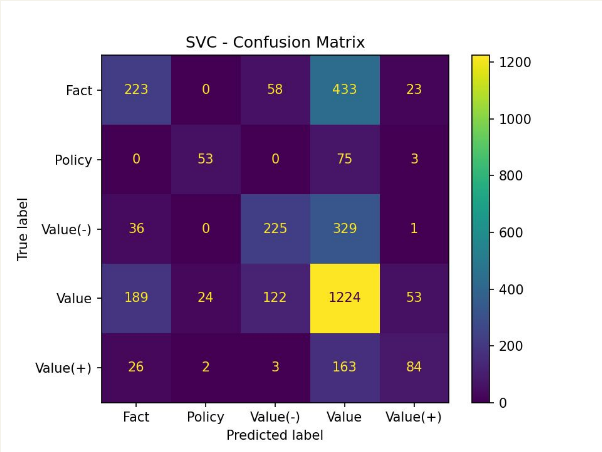


Fig. 4 - Confusion matrix for SVC on unigram and bigram input

OvR ROC-AUC Curves

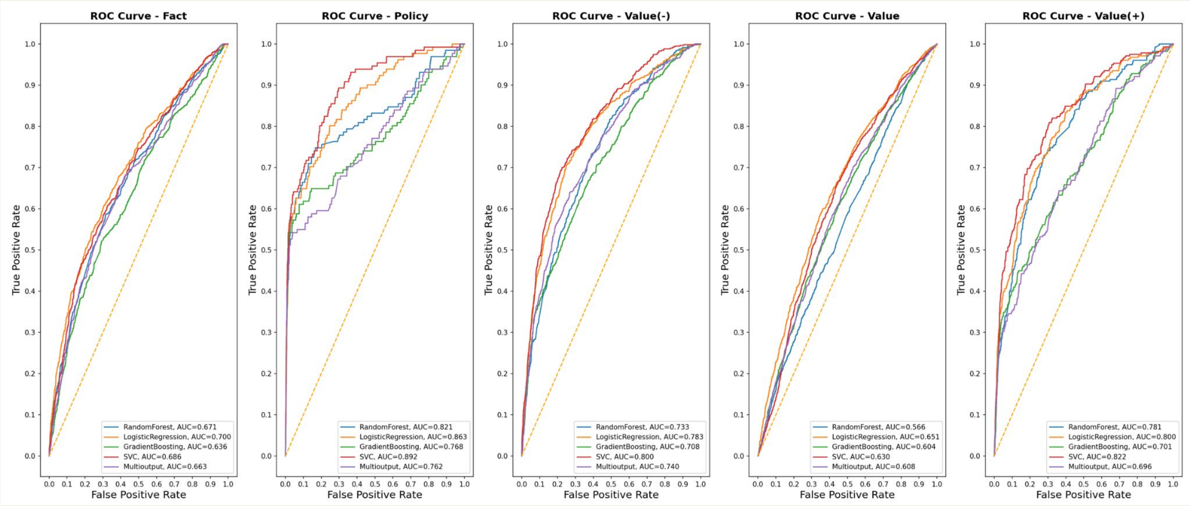


Fig. 5 - One vs Rest ROC-Curves between various models on unigram and bigram input.

07

Conclusions

What were the conclusions of the project? And what could have been done more?

Language

Models for Portuguese language are underdeveloped compared to English, which proved to be a difficulty for processing data

Unbalanced Data

The *Value* class is heavily more present compared to others, which tilted results into defaulting their predictions into that class

Unused Data

In articles dataset, the column containing the keywords of the article could be used to further result performance

Word Embeddings

This representation technique wasn't properly explored and could help improve the results

Linguistic Constructions

Further explore into detection and evaluation of linguistic constructs to detect patterns on each type of ADU

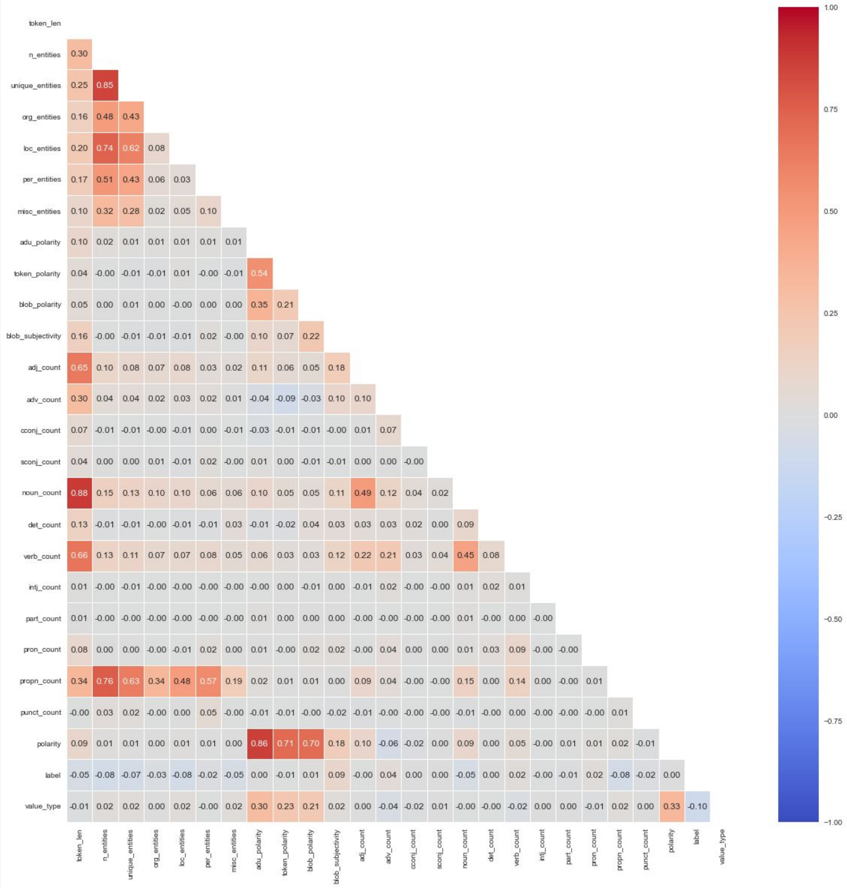
Thanks!

Do you have any questions?

Annexes

Extra resources

Fig.6 - Correlation matrix of extra features to target columns



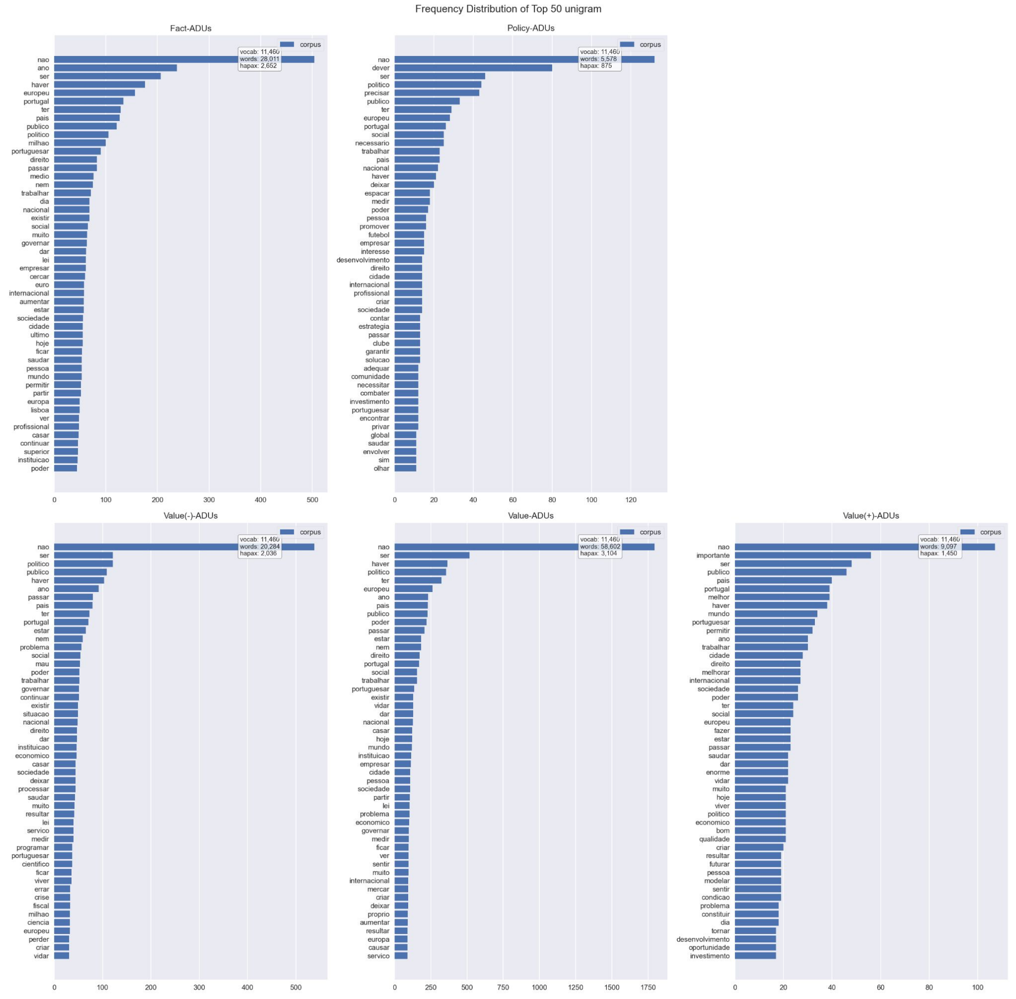


Fig.7 - Count frequency of unigrams for each class of ADU

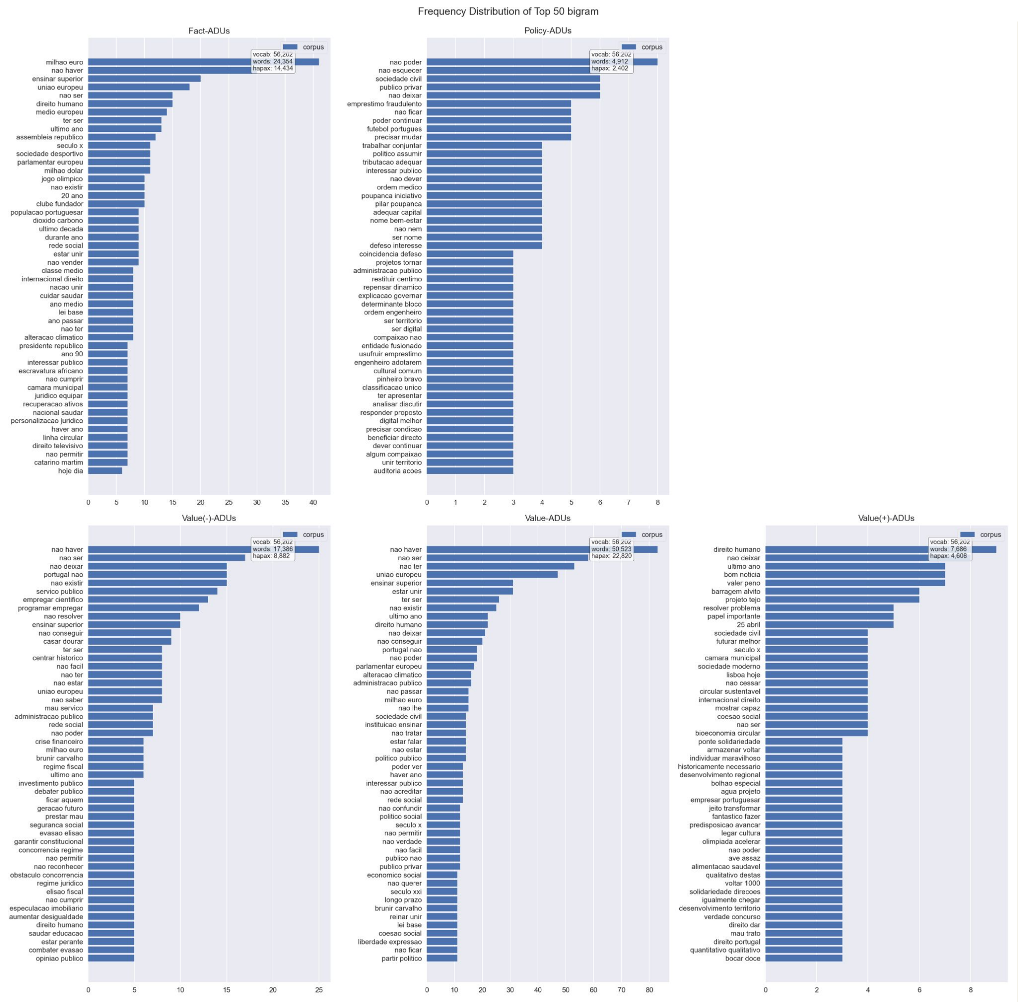


Fig.8 - Count frequency of bigrams for each class of ADU