Discourse Analysis and Argumentation Mining from Text Sources

Inês Quarteu (up201806279@edu.fe.up.pt)

Rúben Almeida (up201704618@edu.fe.up.pt)

20-04-2022

FEUP-PLN 2021/2022

Roadmap

- Introduction
- Dataset
- Preprocessing
- Data Augmentation
- Classification Task
- Representation Techniques
- Algorithms && Parameters
- Results

Introduction

Create ML approach to automatically classify the propositional content of opinion articles segments in Portuguese Language.

Project Word Cloud:

Dataset

Two Files:

- OpArticles.xlsx -> 373 Opinion articles extracted from the Portuguese newspaper Público.
- OpArticles_ADUs.xlsx -> 16743 Argumentative Discourse Units (ADUs) classified by four annotators using as source OpArticles.xlsx

Labels:

- Fact: A statement largely believes its truth or falsehood is widely accepted.
- Value: Statement with an intrinsic evaluative judgment.
- Value(+): Positive Value judgment.
- Value(-): Negative Value judgment.
- Policy: A proportion that advocates a course of action.

Disagreements Between Anotattors:

- For the same ADUs often Anotattors disagree between them in the classification made
- 1023 ADUs have disagreements with 1384 items.
- The level of disagreement varies, ranging from 1vs1 to 1vs4

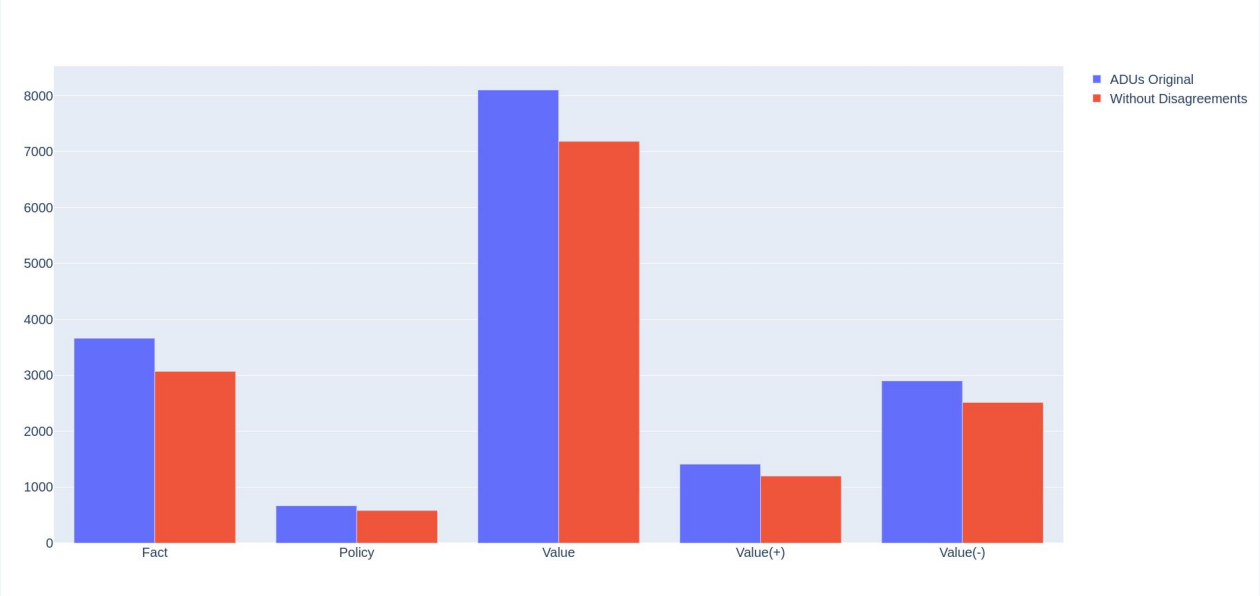
Final Content:

- Discard OpArticles.xlsx content.
- Exclusive usage of the “tokens” field of OpArticles_ADUs.xlsx file

Pre Processing

Dealing with Disagreements:

- Removal: Removing from the dataset all instances where annotators disagree with the label of a range of text
- Majority vote: Find the label most annotators agree on, and remove the from the dataset all others
- Better results were achieved with removal of collisions



Text Processing:

- Tokenization and removal of stop words
- Stemming with SnowballStemmer vs Lemmatization
- Removal of punctuation

Pre Processing

Expanding tokens:

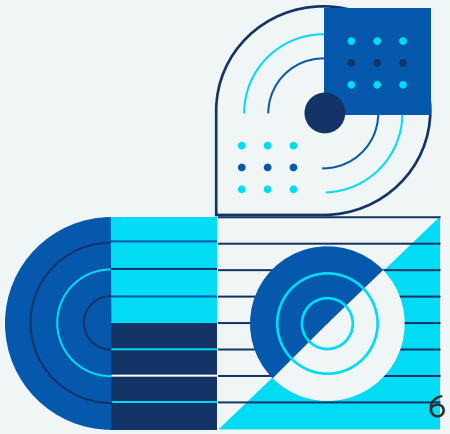
- Adding to each tokens' row the paragraph in which they occur, in order to expand the vocabulary
- Produced a lot of duplicates

POS Tagging:

- Tagged adjectives, interjections, verbs and proper nouns
- Had a positive impact on the model

Lexicons:

- There was no significant improvement with the this technique
- It is hard to find a good lexicon for portuguese
- Many unknown words



Data Augmentation - Key Factor

Motivation:

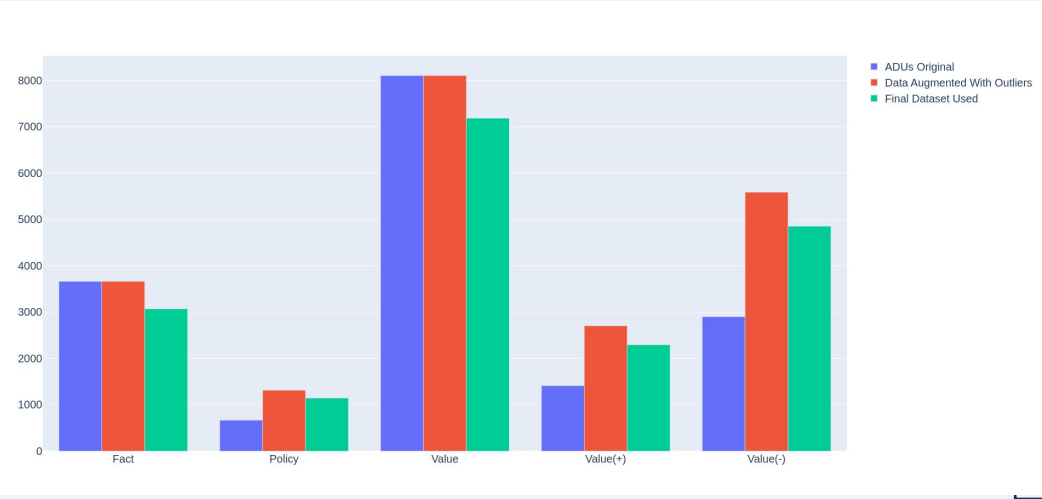
- There was a remarkable discrepancy between the volume of rows labelled as Fact or Value and Policy, Value(+) or Value(-)
- The model does not have a sufficiently expansive set of information to register good results in the last group of labels

Process:

- The tokens of the target labels underwent a multi-layered translation process
- The final tokens that did not represent a duplicate were added to the dataset

Results:

- The targeted labels presented much better results
- This was possibly the studied measure of the greatest impact
- The possibility of overfitting was not addressed



Classification Task

Production Classification Pipeline:

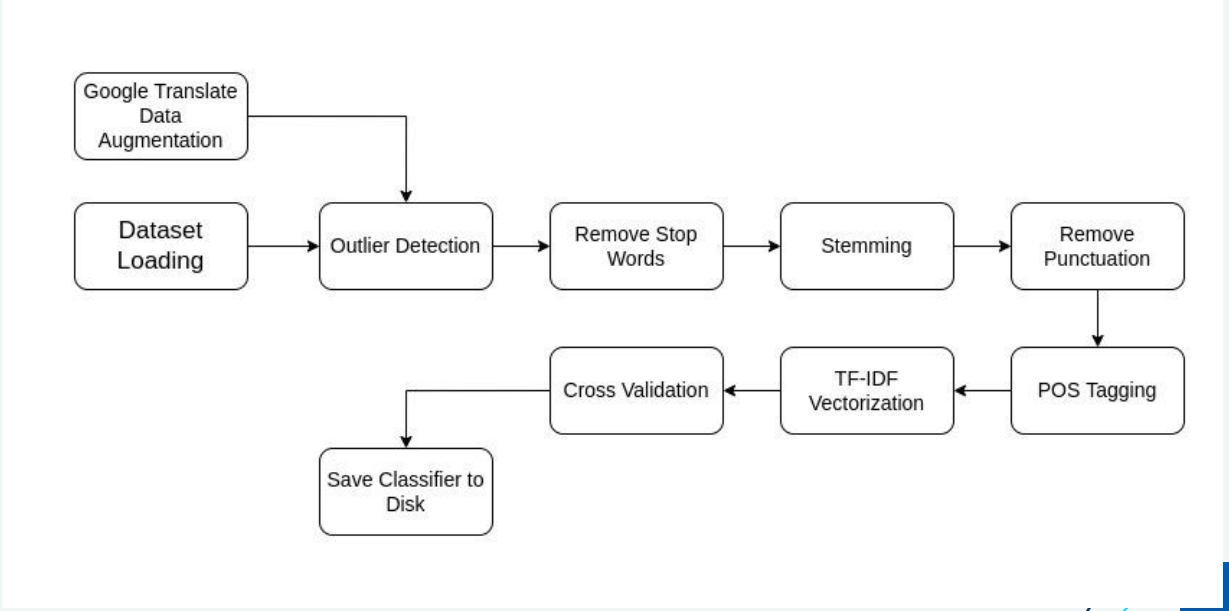
- Exclusive usage of ADUs textual representation
- Discarded biased information based on author or article topics.

Train-Test Techniques Covered:

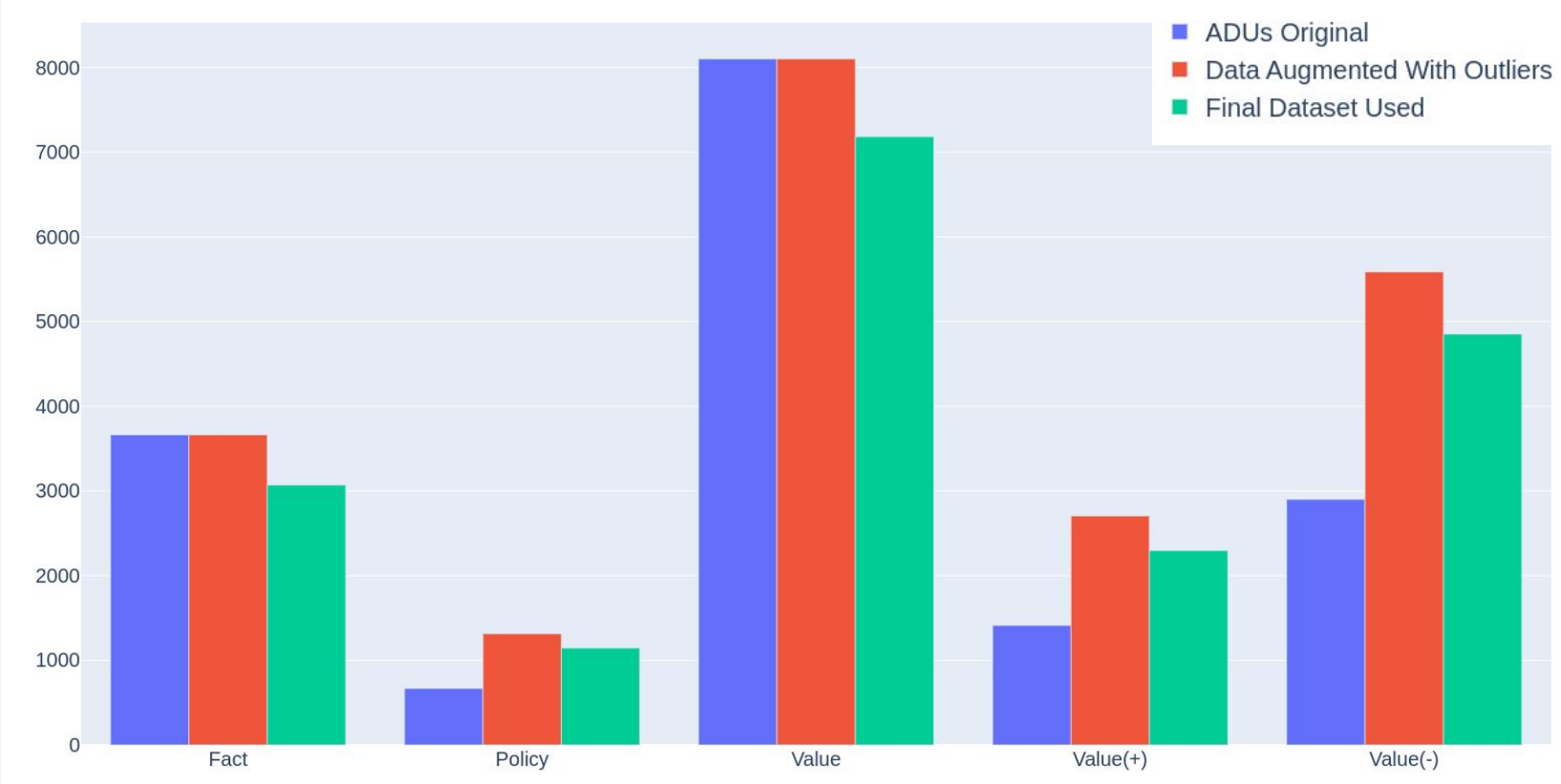
- Randomized Train-Test 80-20%
- Cross-Validation Stratified with 5 folds

Unsuccessful Attempts:

- Train model with the annotator label.
- Train a model for each annotator with the annotator label explicit.
- Train a model for each annotator with the annotator label discarded.



Classification Task - Unbalanced Dataset (II/II)



Representation Techniques

Bag of Word Based:

- CountVectorizer
- One Hot Encoding
- TF-IDF -> Best Results
- Comparison of Unigrams/Bigrams/Trigrams

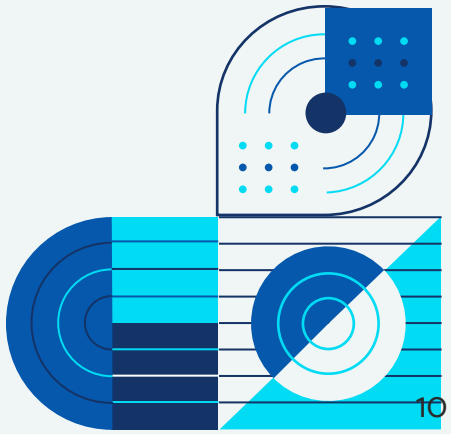
TF-IDF Parameter Tuning:

- Combined with Bigrams
- Vocabulary Size: 41857

Force Lowercase	False
max_df	0.10
min_df	2

Embeddings Based:

- Trained embedding model
- Using a pre-trained model: *skips_100.txt*



Algorithms && Parameters

Best Performing Algorithm:

- SVM
- Sklearn SGDClassifier

SGDClassifier Parameter Tuning:

- Obtained by Grid Search
- Training Time: 3 minutes

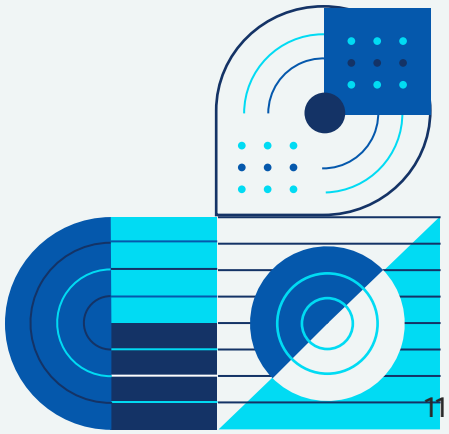
Loss	Modified Huber
Max Iterations	30000
Nº Iteration No Change	20
Class Weight	Balanced

Other Algorithms Tested:

- Naive Bayes -> Used for Baselines
- Knn
- Logistic_regression
- Decision_tree
- Bagging
- LASSO

Algorithms Not Covered due to Hardware Limitations:

- Random Forest
- Neural Network(MLPClassifier)



Results (I/II)

Baseline:

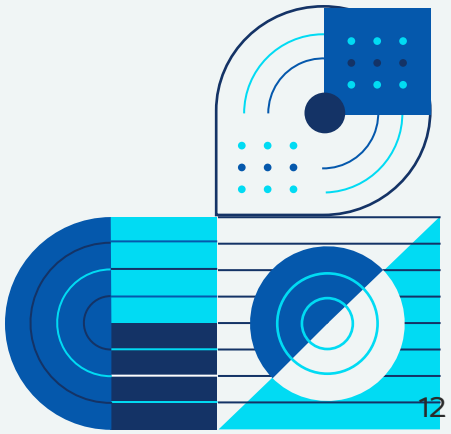
- Naive Bayes
- No Pre Processing
- No Data Augmentation

- AUC ovo Weighted: 0.605
- F1-Score Macro: 0.41

Production Pipeline:

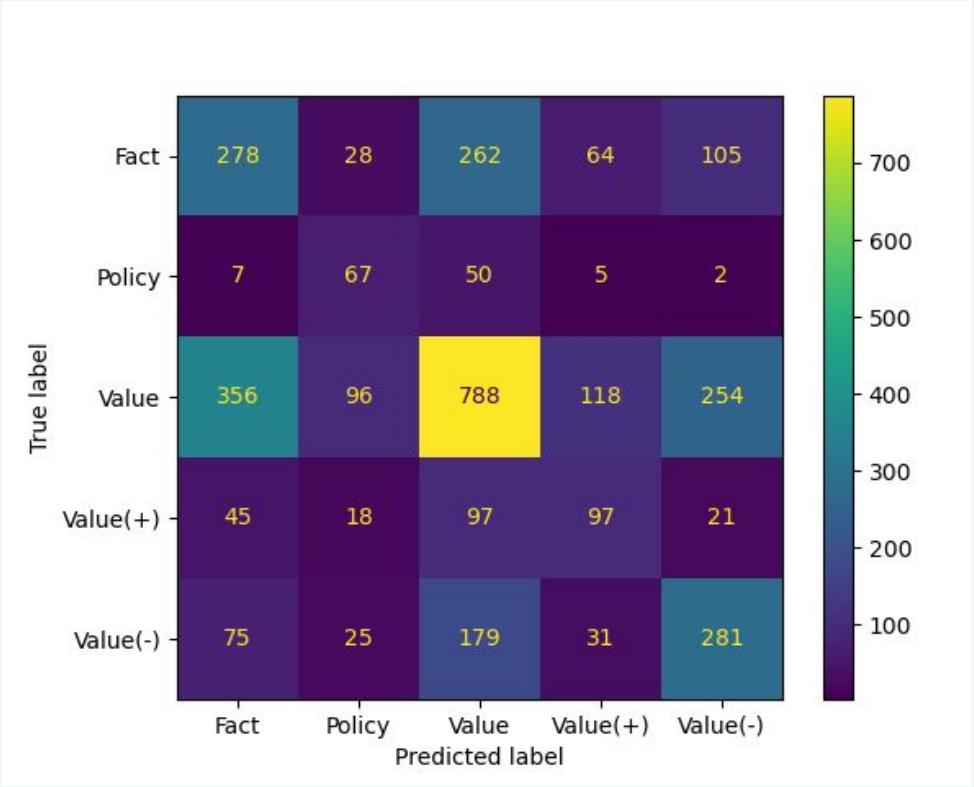
- SGDClassifier
- Loss : Modified Huber
- Class Weight : Balanced

- AUC ovo Weighted: 0.857 (+42%)
- F1-Score Macro: 0.656 (+60%)

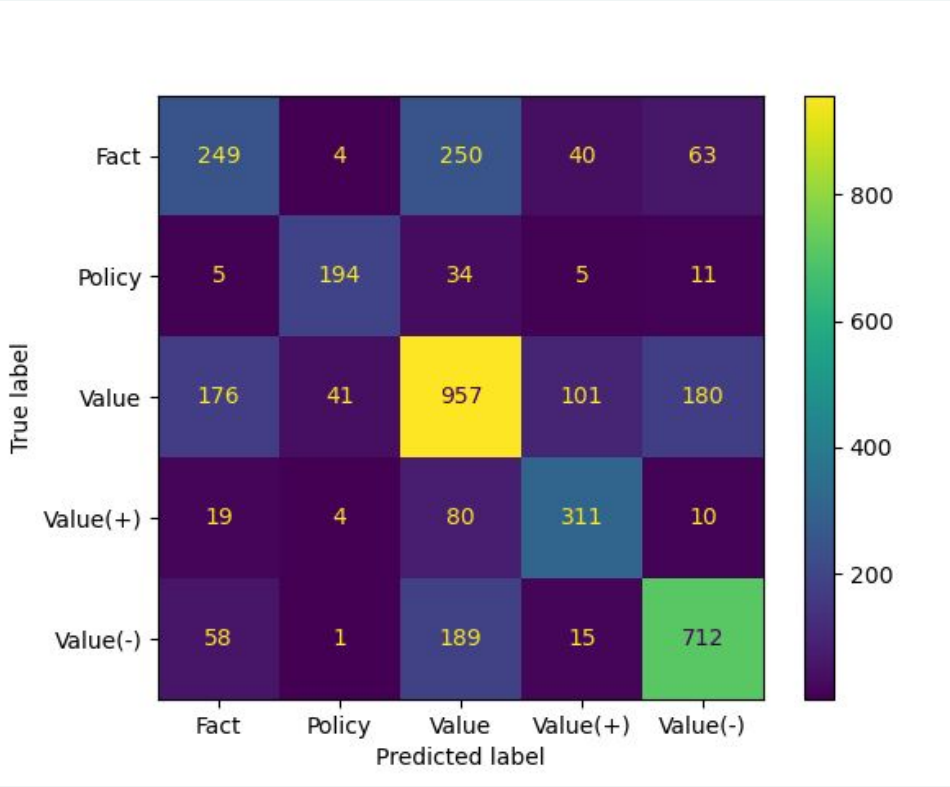


Results (II/II)

Baseline:

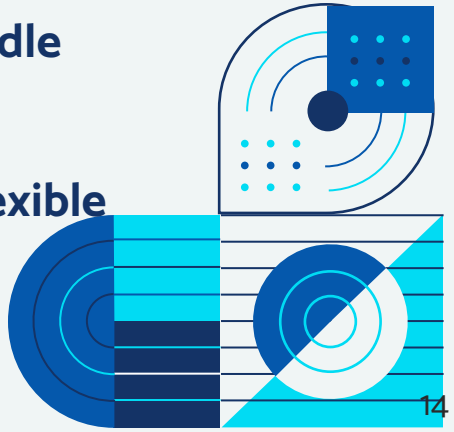


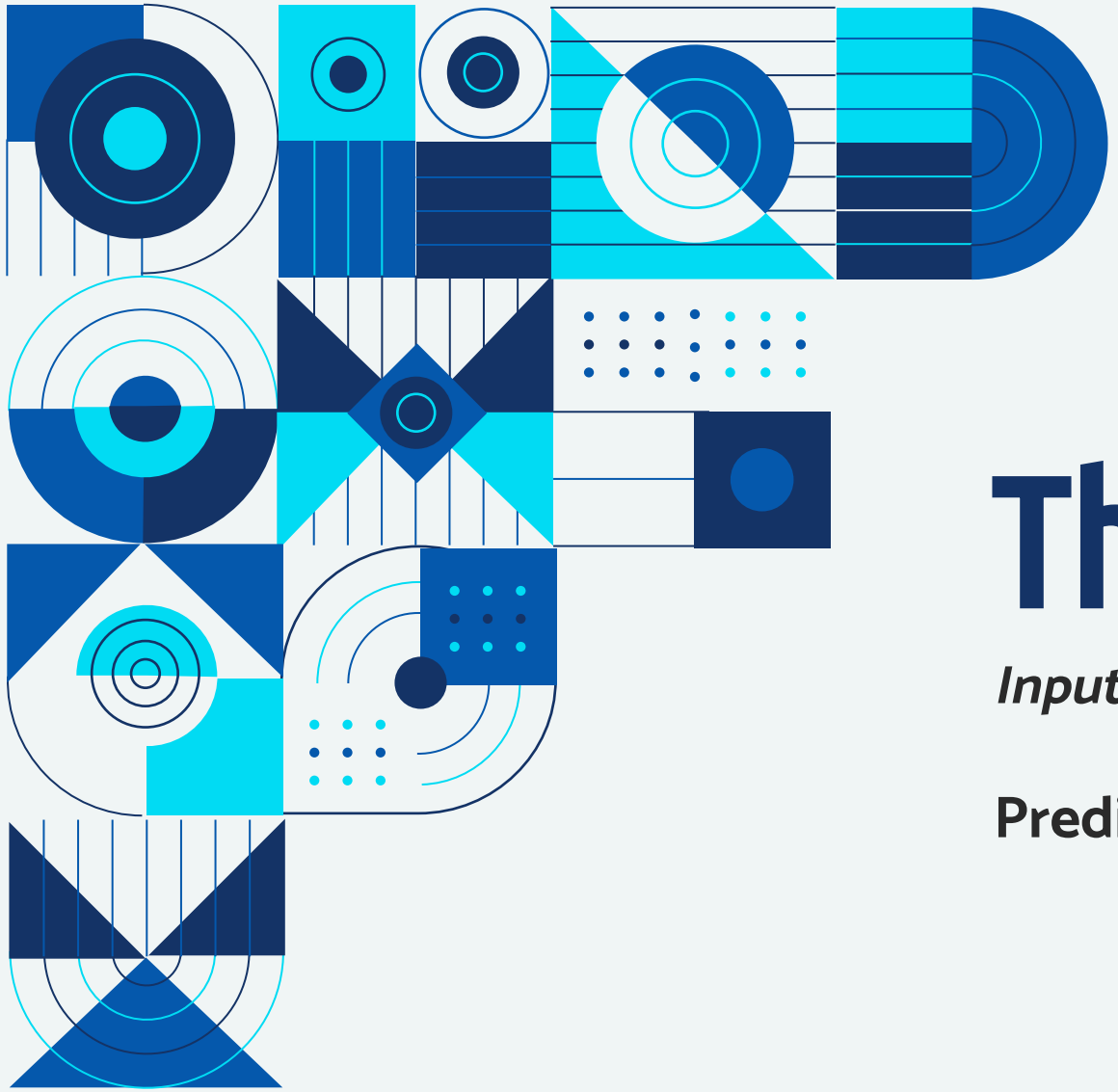
Production Pipeline:



Conclusions

- Many Techniques were applied, the majority without much success or residual improvements.
- Data Augmentation revealed the key factor to escape the low scores.
- SVMs with TF-IDF vectorization is a good combination of techniques to address this classification problem.
- The group was unable to run all ML algorithms, mostly the ones that tend to perform better like R.Forest and Neural Networks. Hardware resources and techniques to handle this kind of situation are required for future developments.
- The final result achieved the goals planned by the group. The final release offers a flexible and consistent pipeline approach to address this ML problem.





Thank you!

Input: Do you have any questions?

Prediction: Value(+)