University of Warsaw

Natural Language Processing

Unfair clauses

Adam Nowak (id: 467298)

WARSAW, JANUARY 2025

1. Introduction

The goal of the project was to develop a program that analyses a dataset to determine whether a given clause is legal or illegal. All relevant files are attached on GitHub: *https://github.com/Adrint/NLP*.

2. Dataset

The dataset used in the project is available at: *https://huggingface.co/datasets/laugustyniak/abusive-clauses-pl*. It consists of training, testing, and validation sets.

- The training set was used to train the model.
- The testing set was used to evaluate the model's performance.
- The validation set was not used in this project.

The training dataset contains two columns:

- text: includes clauses analysed during the project.
- label: identifies the clauses as legal or illegal.

Dataset statistics:

	Legal	Illegal
Number	2338	1946
Percent	55%	45%

Table 1

Example of an original clause:

„W przypadku opóźnień Pożyczkobiorcy w spłacie pożyczki Pożyczkodawca ma prawo raz dziennie telefonicznie upomnieć Pożyczkobiorcę do spłaty pożyczki. Za każdy telefon do Pożyczkobiorcy Pożyczkodawca nalicza opłatę w wysokości 100”

3. Normalizing text

The training set was normalized as follows:

- Removed punctuation marks.
- Converted all uppercase letters to lowercase.
- Removed numbers.

Example of normalized text:

„w przypadku opóźnień pożyczkobiorcy w spłacie pożyczki pożyczkodawca ma prawo raz dziennie telefonicznie upomnieć pożyczkobiorcę do spłaty pożyczki za każdy telefon do pożyczkobiorcy pożyczkodawca nalicza opłatę w wysokości”

4. Preparing data for machine learning

To convert the text into numerical representation, the CountVectorizer class from the scikit-learn library was used:

- Polish stop words (low-importance words) were ignored.
- A dictionary of features was created from the training texts and converted into a numerical matrix.

The testing set was only transformed into a numerical matrix using the previously created feature dictionary.

5. Machine learning

The following algorithms from the scikit-learn library were implemented:

- Logistic Regression.
- Naive Bayes.
- Decision Tree.
- Random Forest.

The models were trained on the training set and tested on the test set. Their performance was evaluated using the following metrics:

- Accuracy: overall correctness of the model.
- Precision: correctness of positive predictions.
- Recall: ability to detect all relevant instances.
- F1-score: balance between precision and recall.

Additionally, confusion matrices were visualized as heatmaps. The results of each model were summarized and compared in a table.

6. Summary

Four models were tested: Logistic Regression, Naive Bayes, Random Forest, and Decision Tree. The best-performing models were Logistic Regression and Random Forest, achieving the highest accuracy (79.8%) and F1-scores (0.85 and 0.849).

- Logistic Regression performed the best at correctly classifying legal clauses (1979) and achieved the highest recall (84.8%).
- Naive Bayes had the highest precision (86.4%) and correctly classified the most illegal clauses (828).
- The weakest model was the Decision Tree, with an accuracy of 76.5%

The results of the models are presented in the table and confusion matrix heatmaps, available on the following page.

The program was delivered as a main.py file along with a file containing Polish stopwords.

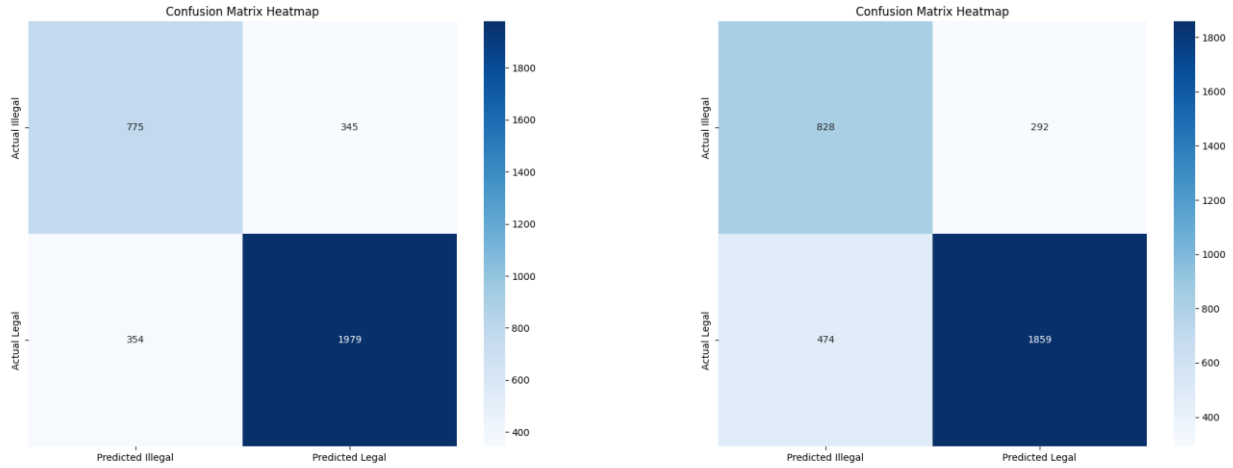

Figure 1 Logistic Regression and Naïve Bayes

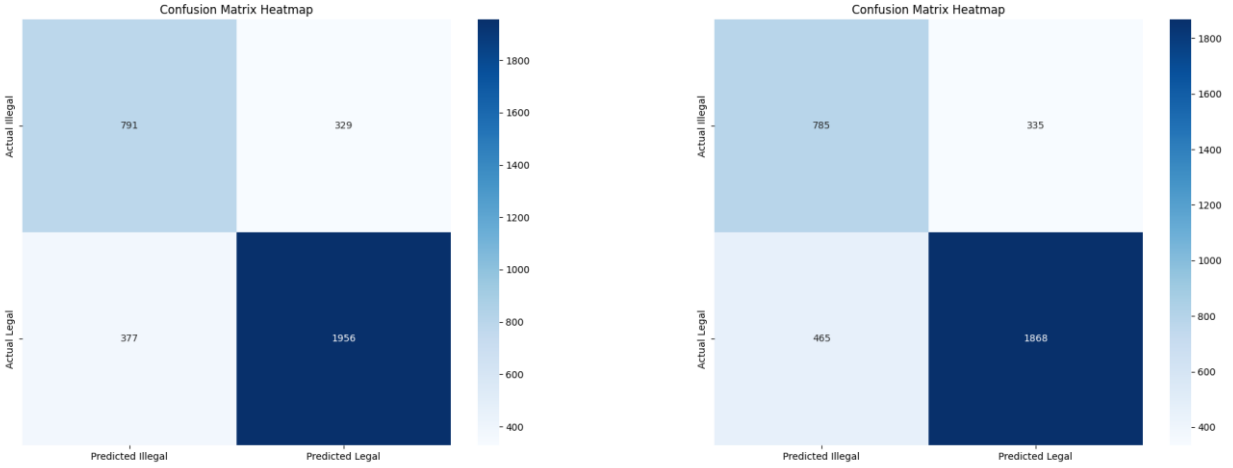

Figure 2 Random forest and Decision tree

	Logistic Regression	Naive Bayes	Random forest	Decision tree
Accuracy	0.798	0.778	0.798	0.765
Precision	0.852	0.864	0.860	0.847
Recall	0.848	0.797	0.838	0.796
F1 Score	0.850	0.829	0.849	0.821

Table 2

	Logistic Regression	Naive Bayes	Random forest	Decision tree
Actual legal and predicted legal	1979	1859	1938	1848
Actual illegal and predicted illegal	775	828	821	791
Actual legal and predicted illegal	354	474	395	485
Actual illegal and predicted legal	345	292	308	329

Table 3