NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "Data Science and Business Analytics"

Software Team Project Report on the Topic:

Research of Criminal Cases and Identification of their Important Characteristics

Submitted by the Student:

group #БПАД222, 3rd year of study	Morozova Yulia Dmitrievna
group #БПАД222, 3th year of study	Sheredeko Arina Yanovna
group #БПМИ2211, 3th year of study	Khugaeva Dana Arturovna

Approved by the Project Supervisor:

Munerman Ilya Viktorovich
Visisting teacher
Faculty of Computer Science, HSE University

Moscow 2025

Contents

Annotation **4**

Keywords **4**

1 Introduction **5**

1.1 General Idea 5

1.2 Tasks distribution 5

1.2.1 Data 5

1.2.2 Characteristics 6

1.2.3 Feature extraction 6

1.2.4 Visualization and ML analysis 6

1.2.5 Other work 6

2 Literature review **6**

3 Made work and results **8**

3.1 Data Collection 8

3.2 Dataset preparation 8

3.3 Characteristics 9

3.4 Manual marking of cases and creation of a "gold" dataset 11

4 Feature extraction **12**

4.1 Extracting characteristics from sentencing texts using LLM 12

4.2 Find numeric characteristics using Regular Expressions 13

4.3 Results summary 14

4.4 Combining results in one new dataset 15

5 Visualisation and analysis **15**

5.1 Creating extra features 15

5.1.1 Distribution of cases by region. 15

5.1.2 Distribution of genders. 16

5.2 Extra visualization using Tableau 17

5.2.1 Descriptive analysis of result types and drug weights 17

5.2.2 Distribution of types of punishments depending on the drug amount 18

5.2.3 Distribution of the drug weights among the cases 19

5.3 Hypotheses Testing 20
5.3.1 Hypothesis 1 20
5.3.2 Hypothesis 2 21
5.3.3 Hypothesis 3 21

6 Vector of future development **22**
6.1 Expanding the training sample 22
6.2 Improving the quality of complex feature extraction 23
6.3 Expanding analytical capabilities 23
6.4 Multifactorial forecast of court decisions 23

7 Conclusion **24**

References **25**

Annotation

This course work examines methods of automated analysis of criminal cases classified under Article 228 of the Criminal Code of the Russian Federation using machine learning (ML) and large language models (LLM). The purpose of the study is to develop approaches to extract key characteristics from court and pre-trial materials without the need for manual processing.In addition to extracting information, the possibility of identifying factors influencing the final court decision is being investigated. To achieve these goals, a representative sample of real judicial texts was formed, preliminary data processing was carried out, and algorithms for structuring and analyzing the contents of documents were implemented.

Аннотация

В данной курсовой работе изучаются методы автоматизированного анализа уголовных дел, квалифицированных по статье 228 Уголовного кодекса Российской Федерации, с использованием машинного обучения (ML) и больших языковых моделей (LLM). Целью исследования является разработка подходов к извлечению ключевых характеристик из судебных и досудебных материалов без необходимости ручной обработки. Помимо извлечения информации, изучается возможность выявления факторов, влияющих на окончательное судебное решение. Для достижения поставленных целей была сформирована репрезентативная выборка реальных судебных дел, проведена предварительная обработка данных и внедрены алгоритмы структурирования и анализа содержания документов.

Keywords

Parsing, Data processing, Text preprocessing, NLP, ML, Large Language Models (LLM)

1 Introduction

1.1 General Idea

Judicial decisions on criminal cases in Russia are an important source for analyzing law enforcement practice, evaluating the effectiveness of legislation, and identifying patterns in the judicial system. However, manual processing of such texts is difficult due to their volume, legal complexity, and heterogeneity of data.

The analysis of cases under Article 228 of the Criminal Code of the Russian Federation, which regulates crimes related to drug trafficking, is particularly relevant. This article covers a wide range of crimes, from storage to sale, and accounts for about 13-15% of all sentences in the country. Drug users are most often prosecuted under it: about 79% of those convicted under part 1 of the article are people without marketing purposes.

In Russia, there is a shortage of empirical research using machine learning and text analysis methods. The lack of annotated open data and methodological contradictions make it difficult to objectively analyze judicial practice. This increases the need for automated approaches capable of extracting key characteristics from texts: mass of substance, region, type of sentence, etc.

In this paper, as a part of a project, we have collected and cleaned up a dataset of real court decisions under Article 228 of the Criminal Code of the Russian Federation. With the help of regular expressions, manual markup, and models based on large language models (LLM), important characteristics were extracted: the type of punishment, the length, the mass of the substance, the presence of mitigating and aggravating circumstances, the gender of the accused, and others. Further, based on the data obtained, we performed visualization, regression analysis, and hypothesis testing to identify possible patterns in the imposition of punishments. The results showed that this approach can be useful for a deeper analysis of judicial practice and the identification of hidden patterns in decision-making.

To get acquainted with the implementation of all tasks, follow the link: project repository

1.2 Tasks distribution

1.2.1 Data

1 Getting dataset. *Arina*

2 Providing primary analysis. *Dana, Yulia*

3 Cleaning data. *Dana*

4 Creating random samples. *Arina*

1.2.2 Characteristics

1 Identification of main characteristic which must be taken from court decision texts manually. *Yulia, Arina, Dana*

2 Manual extraction of important features from court decisions to build a gold standard dataset. *Yulia, Arina, Dana*

1.2.3 Feature extraction

1 Extracting characteristics from sentencing texts using LLM. *Arina*

2 Find integer characteristics using Regular Expressions. *Yulia, Dana*

3 Combining results in one new dataset. *Yulia*

1.2.4 Visualization and ML analysis

1 Creating extra features for visualization. *Yulia*

2 Visualization using Tableau. *Arina, Yulia*

3 A brief interpretation of the results obtained. *Arina, Yulia*

4 Hypothesis Testing. Linear and Logistic regression analysis. *Arina, Yulia*

1.2.5 Other work

1 Writing documentation. *Yulia, Arina, Dana*

2 Creating and filling repository. *Yulia, Arina, Dana*

2 Literature review

Predicting judicial decisions using machine learning and natural language processing has become a prominent area of research in recent years. Early works such as [1] demonstrated that it is possible to predict decisions of the European Court of Human Rights with around 79% accuracy using textual features and support vector machines. This approach was further generalized in the

United States context by [3], who developed a model for forecasting outcomes of the U.S. Supreme Court over decades of case history.

In India, efforts have been made to develop legal decision predictors for the Supreme Court using supervised machine learning. The *eLegPredict* system, trained on over 3000 cases, achieved an F1-score of approximately 76% using gradient boosting models [5].

Recent developments have also incorporated large language models (LLMs) like GPT-4 into the legal domain. [7] evaluated the use of GPT-4 in extracting structured information and predicting outcomes from more than 14,000 UK Employment Tribunal cases. They showed that transformer models, including GPT-4, performed comparably to fine-tuned custom classifiers.

Beyond prediction, generative AI is also used in legal design and education. For example, [4] applied rapid prototyping using LLMs in a legal design course to develop visualizations, chatbots, and service maps, demonstrating how AI can facilitate both legal innovation and student engagement in legal tech.

Several studies have specifically addressed the analysis of legal texts. For instance, [6] proposed an approach to automated legal text analysis using machine learning methods such as TF-IDF and word embeddings, combined with classifiers like Random Forest and SVM. Their work demonstrated high accuracy in extracting facts such as laws, crime circumstances, and penalties. However, the model was built for English-language texts and did not account for domain-specific elements such as drug-related offenses.

Addressing the Russian legal system directly, [2] explored the use of natural language processing for extracting structured data from Russian court decisions. Using the Natasha and SpaCy libraries, they successfully extracted key legal elements such as Criminal Code articles and penalties. Nonetheless, the study employed outdated NLP tools and did not tackle the complexity of drug-related cases or integrate modern deep learning approaches or LLMs.

Overall, these studies illustrate the potential of combining legal data with machine learning and LLMs to support judicial decision-making, legal research, and public access to justice. However, limitations such as language coverage, outdated tools, or generic extraction pipelines point to the need for specialized systems, particularly for sensitive domains in Russian court practice. Our project aims to bridge this gap using modern tools, fine-tuned on domain-specific Russian texts to extract features.

3 Made work and results

3.1 Data Collection

To collect and prepare the corpus of court decisions related to Article 228 of the Criminal Code of the Russian Federation, a specialized parser "Если быть точным" was used — an open-source tool recommended by the supervisor. This parser collects structured data from the official websites of Russian courts, extracting both meta information (case number, court, region, date) and the full texts of sentences. A special feature of the parser is the ability to filter by specific article and type of cases, which made it possible to immediately limit the sample to only criminal cases containing article 228.

To start we choose specific filters:

- Type of case: Criminal Code
- Article: 228
- Format: CSV
- Court level: district, regional
- Region: any
- Years: 2015-2024

Data collection required periodic manual CAPTCHA checks, which is due to restrictions on the part of court portals.

3.2 Dataset preparation

- Volume: 7.16 GB
- Number of rows (cases): 303,783
- Number of columns 25
- The main fields include: `id`, `entry_date`, `result_date`, `in_favorites`, `court`, `codex_articles`, `defendants_simple`, `penalty_type`, `result_text`, `result_text_url`, `court_city`, `region`, `type`, `stage`, `url`, `appeal_date`, `defendants_gender`, `judge`, `case_number`, `last_scheduled_hearing_date`, `in_favorites.1`, `sm_url`.
- The original 7.16 GB dataset was reduced to 223 MB by removing empty columns, links, and incorrect entries using the `dropna()` function and creating random sample from the obtained values.

3.3 Characteristics

During the course of the project, key characteristics to be extracted from the court rulings were identified and selected:

- **result_type:** String, the type of punishment that has been imposed under the sentence. It can take one of 6 values: “штраф”, “условное осуждение”, “лишение свободы”, “ограничение свободы”, “обязательные работы”, “исправительные работы”. This variable is the main target variable for analysis. **Search context**: is searched for at the end of the text, after the words: “приговорил:”, “п р и г о в о р и л:”, “ПРИГОВОРИЛ:”. Next to this characteristic there are expressions like: “назначить наказание в виде”, “приговорить к”, “наказать в виде” etc.

- **sentence_years:** Int, the number of years of sentence imposed by the court (e.g., imprisonment, restriction of freedom, correctional labor).

- **sentence_months:** Int, the number of months of the basic sentence. Used to specify the term if it was imposed in months or in combination with years.

- **sentence_days:** Int, the number of days of the sentence, used if the term is expressed in days (for example, in case of “обязательные работы” or short terms of arrest). It is rare in practice, but is recorded when available.

- **sentence_hours:** Int, the number of hours of compulsory labor. This sentence is used when the court imposes community service in hours. It is important for analyzing light sentences.

- **is_suspended:** Bool, binary feature that reflects whether conditional punishment is imposed. Values: 1 - the punishment is conditional, 0 - the punishment is not conditional. This characteristic that allows to separate the severity of sentences. **Search context**: Next to this characteristic there are expressions like: “условно”, “условное осуждение”, “наказание считать условным” etc.

- **suspention_years:** Int, number of years of suspended sentence, i.e. the term during which the convicted person must not commit new crimes in order to avoid the real execution of punishment.

- **suspention_months:** Int, number of months of suspention, used in addition to suspention_years to more accurately represent the length of the suspended sentence.

- **fine_is_fixed:** Bool, It is a binary feature that reflects whether the amount is fixed in monetary terms (for example, "20,000 rubles") or whether it is set as a share of income ("50% of salary for 3 months"). This metric is important since fixed amounts are are more common in cases with certain social background, while shared amounts are more rare. **Search context**: keywords “штраф”, “в размере”, “в сумме”, “в доле”, “от дохода”. So, while using needed models algorithm analyzes neighboring words and sentences structure to correctly classify the type.

- **fine_amount:** Int, The amount of the fine in rubles specified in the verdict. It can be represented as a single digit ("100,000 rubles"), or with a transcript ("100,000 (one hundred thousand) rubles"). **Search context**: “штраф”, “в размере”, “оплатить”, “назначить” and next to it are numeric values + the word “rubles”.

- **drug_weight:** Float, One of the most significant feature. It is indicated in grams, milligrams, or kilograms. The mass can be represented in different forms — “вещество массой 1,34 г”, “весом 950 миллиграмм”, “высушенный остаток — 0.52 г“. It is also possible that several different substances may be present in the same case.

- **drug_type:** String, This feature is the name of the drug appearing in the text of the verdict. It is important to analyze which substances are most often seized, how they are distributed by region, as well as to conduct a cross-analysis with other characteristics (for example, weight or type of punishment). The type of substance is an important factor in law enforcement, since legislation provides for different thresholds for determining "significant", "large" and "especially large" size for different drugs. **Search context:** The feature can be extracted based on keywords that usually indicate the substance: “наркотическое средство — [название]”, «обнаружено вещество: [название]”, “изъято [название]”, “содержит [название наркотика]”, “вещество, содержащее [название]”. It is also important to create a dictionary and a dictionary of synonyms: “марихуана” → “конопля”, “анаша”, “каннабис” “спайс” → “синтетические каннабиноиды”, “курительные смеси” “героин” → “диацетилморфин”

- **drug_purpose:** String, The characteristic defining the intended purpose: "personal consumption", "sale", "transportation", "manufacture", etc. The context is identified on the basis of key phrases: “с целью сбыта“, “для личного пользования“, “транспортировка наркотиков“. In cases where the target is not explicitly specified, the "unknown" category is used.

- **aggravating_factors:** String, It is a multiple textual feature containing aggravating circumstances, if they are mentioned in the verdict. Usually, they are listed in the final part of the text, starting with the phrase “aggravating circumstances are ...”. Key cases include recidivism, crimes within a group of persons, especially large size, etc.

- **mitigating_factors:** String, The factor is similar to the previous one, but with mitigating circumstances. It is searched for phrases such as: “смягчающими обстоятельствами можно считать”, “в качестве смягчающего обстоятельства”, “имеет на иждивении”. Among the frequently occurring factors are admission of guilt, having children, positive characteristics, self—incrimination, assistance to the investigation, etc.

- **drug_amount:** String, This feature reflects the legally significant classification of the volume of the substance appearing in the case: “значительный”, “крупный”, “особо крупный” размер. Such categories have a direct impact on the qualification of the crime and the severity of the sentence, as they are directly related to the part of article 228 under which the accused is tried (for example, part 1 — for significant, part 2 — for large, etc.).

- **location:** Additional data, namely two columns with latitude and longitude for all administrative districts in the dataset for further complete visualization.

3.4 Manual marking of cases and creation of a "gold" dataset

In order to check how accurately our models for extracting features from court decisions work, we first assembled a reference set of cases marked up manually. Each team member took on 50 sentences, resulting in a total of 150 cases in which we manually identified all the important characteristics.

During the markup process, key characteristics were manually extracted from each sentence text, which were previously approved as significant for analysis. The markup was carried out in Google Sheets and then the data was saved in CSV format.

This file has become the basis for evaluating the quality of models: we ran our regulars and models on new texts and compared the result with what we had marked up manually.

This step helped us to check the accuracy for each feature and look which characteristics the models predict well, and which ones need to be improved.

4 Feature extraction

4.1 Extracting characteristics from sentencing texts using LLM

The goal was to automatically extract the **result_type** characteristic from the full text of court sentences using a **Large Language Model**. This is an important step in automating the analysis of court documents, allowing for faster statistical and legal analysis.
The `DeepPavlovrubert-base-cased` model, a pre-trained BERT-based transformer model adapted for Russian, was chosen for the task.

To prepare court decision texts for prediction of the `result_type`, the key part was extracted from the sentencing texts, from the word "приговорил:" onwards. The ruBERT model was pre-trained on our manually labeled dataset for the task of classification into 6 types of punishments (discussed above in `result_type` characteristic). It was applied to a dataset of 10,000 rows divided into chunks of 1,000 rows each. The model achieved accuracy of 0.9463 on our manually labeled dataset. Individual classes reached F1 to almost 1.00 (e.g., "штраф", "условное осуждение", "лишение свободы".

Challenges and difficulties: Training the model on 150 lines is a limitation, but the model showed high accuracy. Texts are usually very long, so we had to truncate texts to 512 tokens, which could affect the completeness of information extraction. Also, the `result_type` could be formulated in different ways. That is the main reason we have cut the most promising part of text and used the LLM for this text classification task. We have tried different pipelines like feature extraction or text generation, but neither of them worked well for our case.

Also we used the same model and pipeline for the extraction of `drug_amount` feature. A key part of the sentence text was extracted using regular expressions, the part that mentions the amount of drug. Weighted losses were used to compensate for class imbalance. Individual classes reached F1 to 0.94 and 0.78 for "значительный", "крупный" respectfully.

Challenges and difficulties: The model cannot predict "особо крупный" class because it is missing from the training. Training on 150 rows is a limitation, but the results are still acceptable. Text can describe the amount of a substance in a wide variety of contexts. Therefore, only the key part of the text was extracted.

Additionally, we attempted to automatically extract the **mitigating_factors** feature from

the sentencing texts. This characteristic could contain multiple meanings, such as “признание вины”, “наличие иждивенцев”, and so on. This fundamentally differentiated the task from previous ones, as it required *multi-valued classification* rather than simple one-class prediction.

For this task, we used the same `DeepPavlov/rubert-base-cased` model, but tuned it for multi-attribute inference. We pre-defined a list of 21 possible mitigating circumstances and manually annotated a dataset of 150 examples, where each sentence corresponded to a list of applicable mitigating circumstances. The model was configured to produce a binary vector of 21 items indicating the presence or absence of each factor.

Problems and challenges: Despite using the same LLM-based pipeline that has worked well for other features, this task proved to be much more difficult. Firstly, the training dataset was very unbalanced: most examples contained only the most common factors, such as “признание вины” and “раскаяние”, while most other factors were either underrepresented or missing entirely. As a result, the model is quickly reconstructed and always predicts only the two most frequently occurring classes. Secondly, due to the small size of the dataset, the model was unable to learn the structure of the task. Unlike `result_type` and `drug_amount`, which can often be identified at a fixed position in the text, mitigating factors are scattered throughout the document and expressed in a wide variety of wording. In many cases, even a human required contextual understanding and legal training to reliably recognize a factor. Therefore, the model had a macro F1-score of only **0.0892**, and accuracy and recall were close to zero for most classes except two most common ones. The model tended to produce the same fixed list of 1-2 factors for each case, regardless of the actual content of the text.

4.2 Find numeric characteristics using Regular Expressions

Regular expressions are a tool for searching and extracting information from unstructured text based on predefined patterns. They are widely used in natural language processing (NLP) tasks because they allow to find certain structures in the text: numbers, dates, names, abbreviations, keywords, etc. Unlike machine learning models, regular expressions do not require training on data, but work on a "find by pattern" basis.

Our teammate *Dana* from another studying program used Regual Expressions to get sentence_years, sentence_months, sentence_days, sentence_hours, is_suspended, suspention_year, suspention_months, fine_is_fixed, fine_amount.

The regular expression model was used to get drug_weight.

To extract the mass of narcotic substances from the texts of sentences, a Python script was implemented using the pandas library for processing tables and the re module for applying regular expressions. The regular expression was configured to search for numbers accompanied by the words "gram", "gr.", "g.", taking into account different forms and abbreviations. Both integer and fractional values separated by periods or commas were also taken into account. Additionally, the filtering function is_mass was used, which made it possible to exclude incorrect values, such as too large numbers or incorrectly extracted fines.

To increase accuracy, the algorithm limited the context analysis — when the first mention of the mass was found, the search ended at the nearest point or line translation. This made it possible to avoid mistakes related to getting other numbers into the analysis area (for example, from medical reports). If there were several acceptable values in the text, they were combined into one line separated by ";". However, this implementation had its limitations: it did not take into account weight in other units (for example, milligrams or kilograms, since it is a rare amount type for drugs in general), did not bring everything to a single format (grams), did not automatically summarize doses, and could skip values if they were outside the selected text fragment.

Difficulties were also caused by non-standard formulations, spaces inside numbers, or cases where the mass was indicated by words ("one gram"). All this reduced the accuracy in some cases. In the future, this task could be improved by using NER models or advanced LLMs that can recognize the mass of a substance in a more flexible and context-sensitive way, as well as automatically convert values to grams and aggregate them. Nevertheless, even the current approach based on regular expressions has shown high practical effectiveness and has become a reliable part of the feature extraction pipeline.

4.3 Results summary

On the Table 4.1 the accuracy scores calculated on our manually labeled dataset of 150 cases are given. As it is shown, the feature extraction models achieve consistently high accuracy across all fields, with most scores exceeding 0.90. Notably, the extraction of critical legal attributes such as `result_type`, `is_suspended`, `suspension_months`, and `fine_amount` demonstrate excellent precision — highlighting the reliability of our RE and NLP pipeline in handling real-world Russian court texts. These results underline the robustness of our methodology and its practical applicability for structured data extraction from unstructured legal documents, particularly in the context of drug-related court cases under Article 228 of the Russian Criminal Code.

Table 4.1: Table of accuracy scores for extracted features

Feature	Accuracy
result_type	0.9463
sentence_year	0.8800
sentence_month	0.9267
sentence_day	0.9133
sentence_hour	0.9467
is_suspended	0.9600
suspension_years	0.9067
suspension_months	0.9800
is_fixed_fine	0.9600
fine_amount	0.9333
drug_weight	0.9400
drug_amount	0.9060

4.4 Combining results in one new dataset

After extracting the features using different approaches (manual markup, regular expressions, LLM), the next step was to combine all the results into a single final dataset. This was necessary for subsequent analysis, visualization, and evaluation of extraction quality.

The first stage was to combine all the original text data on sentences into one file — 10 separate CSV files with cases (from 0 to 10 thousand). Using a Python script, the files were glued together line by line, and the duplicates of the headers were deleted.

Then all the predictions from different models were combined, for example, the values extracted using regular expressions (sentences, fines, conditionality). All files were combined into one dataframe and merged with the main dataset by the id field.

At the last stage, a final check of the structure was carried out, duplicates were removed and it was checked that all the signs were in place. Also, at some points, the necessary signs were added and removed for additional and clearer visualization.

5 Visualisation and analysis

5.1 Creating extra features

5.1.1 Distribution of cases by region.

For in—depth analysis and visual visualization of the data, additional processing of a geographical feature was implemented - the region of the case.
At the first stage, a column with the name of the region was selected from the total dataset (10,000 cases). After that, a dictionary of correspondences was manually compiled, in which each region was mapped to its coordinates — latitude and longitude.

Using the Python function, the latitude and longitude values were added to the main dataset in the form of two new columns: `latitude` and `longitude`. This made it possible to link each case to a specific geographical point on the map of Russia.

Next, the data was grouped by region and the total number of cases in each subject was calculated. The resulting aggregated values were exported to a separate file and used to build an interactive cartogram in Tableau.

The final map (Figure 5.1) shows how criminal cases related to Article 228 of the Criminal Code of the Russian Federation are distributed throughout the country. The color scale reflects the case density: from light green (fewer cases) up to bright red (highest concentration). This visualization made it possible to quickly identify regions with an abnormally high number of cases.

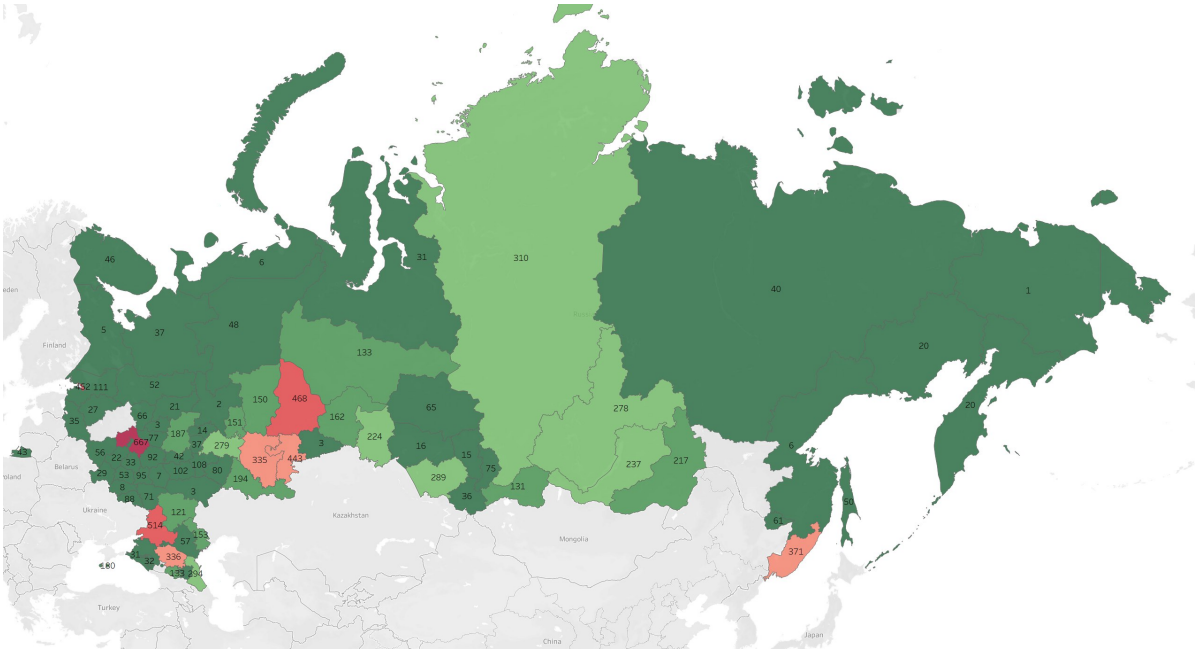


Figure 5.1: Distribution of cases by regions

5.1.2 Distribution of genders.

The analysis of the gender distribution among the accused was carried out on the basis of a purified and normalized `gender_group` attribute. In the source data, the field containing information about the defendants' gender often included lists (in cases with multiple defendants), as well as incomplete or ambiguous entries (for example, "-", "M, F", "M", "F."). For correct

analysis, preprocessing was applied, during which the values were reduced to unified categories:

- Men — only men
- Women — only women
- Mixed — both men and women are present
- Unknown — gender is not specified or cannot be classified

After processing, it became possible to build an adequate gender distribution. The distribution looked like this:

Table 5.1: Table of gender distribution

Gender	**Amount**
Women	777
Men	8480
Mixed	50
Unknown	476

Thus, after clearing the data, it was possible to obtain a representative distribution by sex of the accused. Most cases involve men, women are much less common, and cases with an unspecified or mixed composition are rare. These data highlight the pronounced gender imbalance among the defendants under Article 228 of the Russian Federation Criminal Code.

5.2 Extra visualization using Tableau

5.2.1 Descriptive analysis of result types and drug weights

The top-left plot on the figure 5.2 shows the average log-transformed weight of the substance for each predicted result type. Surprisingly, suspended sentences and corrective labor are associated with the highest average log drug weights (2.08 and 2.00 respectively). Imprisonment appears lower (1.95) and fines much lower (0.59), with freedom restriction being the lowest (0.39).

These averages may be skewed by outliers or reflect confounding factors. For example, there are observations with extra large weights that are somehow predicted to get the suspension sentences. It may be the error during prediction of `result_type` or drug_weight, or it may be the case of corruption in Russian courts.

They challenge the assumption that only higher weight leads to stricter punishment. So, the weight should be analyzed together with other features, not independently.

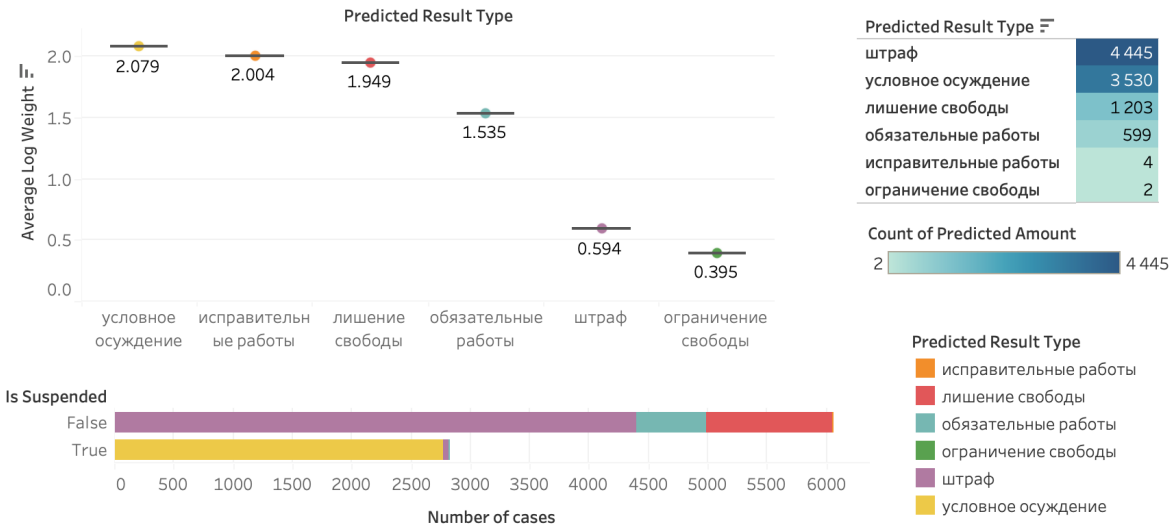


Figure 5.2: Punishment types, drug weights, and suspension status dashboard

The top right table shows the distribution of predicted types of punishment across all cases. The most common type of punishment is “штраф” (4,445 cases), followed by “условное осуждение” (3,530) and “лишение свободы” (1,203). Other types of punishment, such as “обязательные работы”, “исправительные работы” и “ограничение свободы”, are much less common. This indicates that in most cases courts choose less severe non-custodial measures.

The bottom bar chart shows the number of cases where the sentence was suspended or not based on binary `is_suspended` characteristic and the distribution of punishment types. We can see from this graph that there are errors in identification of the `result_type`, but they are not significant. There is a strong relationship between type of punishment and whether the sentence is suspended, which supports the need to analyze this variable explicitly in later models.

5.2.2 Distribution of types of punishments depending on the drug amount

The visualization shown on 5.3 shows the distribution of the types of penalties imposed for two categories of seized narcotic substances: “значительный” и “крупный”, according to the predicted values of the attribute predicted_amount.

For the category of “значительный”, there is a predominance of mild forms of punishment: fines (about 37% of cases) and suspended sentences (about 30%). Actual terms of imprisonment are imposed only in about 10% of cases.

In the “крупный” category, the structure is changing dramatically: the proportion of actual incarceration is increasing, while the frequency of fines and suspended sentences is significantly decreasing.

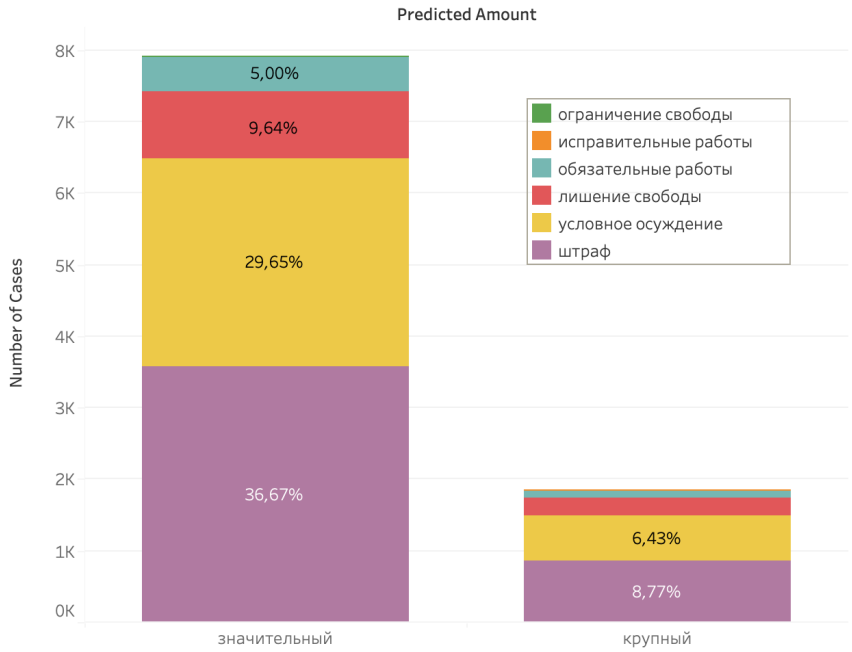
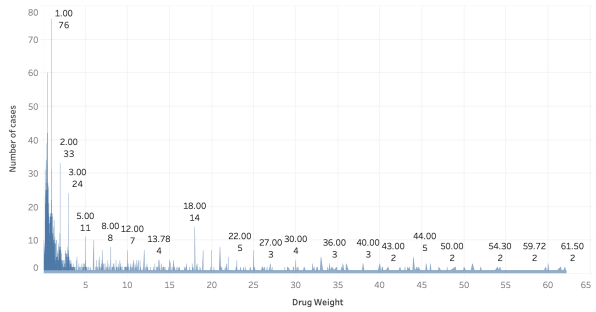


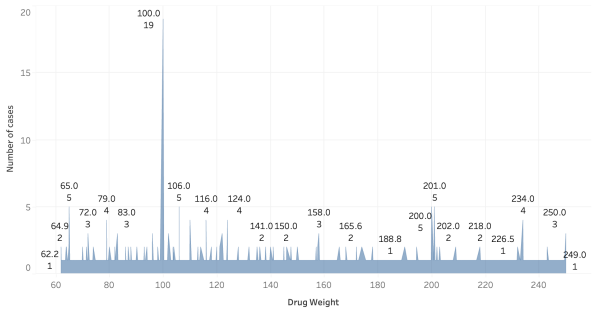
Figure 5.3: Distribution of result_type by drug_amount category

This graph confirms the hypothesis that with an increase in the volume of seized substances, the likelihood of more severe penalties, such as imprisonment, increases significantly. This is consistent with the provisions of the Russian Federation Criminal Code, according to which the size of the substance directly affects the qualification of the act and the sanctions in the case.

5.2.3 Distribution of the drug weights among the cases



(a) from 0 to 62 grams



(b) from 62 to 250 grams

Figure 5.4: Number of cases distribution for drug weight for different scale

These two graphs in Figure 5.4 provide valuable insights about the distribution of the weight of drugs in criminal cases, divided into two ranges: from 0 to 62 grams (left) and from 62 to 250 grams (right). The first graph shows a dramatic accumulation of cases with a weight of about 1 gram (76 cases with precisely this value). This may indicate that 1 gram is the threshold for qualification of the offense, and also indicates the typical volume of the substance

that is confiscated during arrest. In addition, there are peaks at 2 g, 3 g, 5 g, and 18 g which are probably typical "operational" volumes. The second graph shows single spikes at 100 and 200 grams, which is most likely related to the transitions between the categories of "значительный" and "крупный" amounts under Article 228 of the Criminal Code of the Russian Federation. The presence of such clear-cut steps may indicate the influence of purposeful actions of police officers: in Russian practice, there are cases of drop-off of drugs in order to fulfill a plan for solving crimes or to put pressure on specific individuals. Thus, the structure of substance mass distribution may reflect not only the real drug situation, but also manipulative administrative practices that distort the statistics of criminal cases.

5.3 Hypotheses Testing

5.3.1 Hypothesis 1

Larger drug amount increases the likelihood of more severe types of punishment.

We used a **multinomial logistic regression model** to test how a binary indicator of drug quantity (predicted_amount_bin) influences the type of punishment assigned on the significance level of 5%.

Significant finding:

- The only statistically significant effect of predicted_amount_bin = 1 is for the outcome "условное осуждение" (suspended sentence).
- The coefficient is –0.2403, with p-value = 0.004.

That shows us that a **large drug amount significantly reduces the likelihood of receiving a suspended sentence relative to imprisonment**.

Near-significant effect:

- "штраф" (fine).
- Coefficient = –0.1421, p-value = 0.075.

This suggests a **possible decrease in the likelihood of receiving a fine instead of prison for large amounts of drugs**.

For all other punishment types, the influence of amount is not statistically significant. In categories like "ограничение свободы", "принудительные работы", model convergence issues and huge standard errors suggest there are **too few cases** to draw reliable conclusions.

5.3.2 Hypothesis 2

The larger the drug amount leads to the longer the prison sentence, among those who received actual imprisonment.

We filtered the dataset to include only cases with actual imprisonment (predicted_result_type = "лишение свободы"). The independent variable is the logarithm of drug weight (log_weight) was used to normalize the scale and reduce skewness. After that, we regressed sentence length in months (sentence_total_months) on log_weight using linear regression (OLS from statsmodels).

- Coefficient = +2.1545, p-value = 0.023.
- R-squared = 0.004.

The statistical analysis reveals a positive and significant coefficient (p-value = 0.023), demonstrating a correlation between greater drug weights among incarcerated individuals and longer sentences, with an average increase of approximately 2.15 months per unit increase in the logarithmic weight. However, the R-squared value is low (0.004), indicating that drug weight explains only a small portion of the variability in sentence length. While supporting the link between larger drug weights and extended sentences, the model's overall explanatory effectiveness is limited. Drug weight, while statistically significant, is not a strong standalone predictor of sentence length. The reason could be the drug substance severity - not only the weight is important, but the combined information about the weight and the drug substance.

5.3.3 Hypothesis 3

Women receive more lenient punishments than men, without taking into account other variables.

To assess the effect of the defendant's gender on the likelihood of receiving a lenient sentence (fine or suspended sentence), a simplified logistic model was built that includes only the binary variable female as a predictor. According to the regression results **(n = 9257)**, the coefficient for female turned out to be positive **(0.0987)**, which indicates a slightly higher probability of

assigning mild punishment to women compared to men. However, this effect is not statistically significant **($p = 0.316$)**, and therefore, the data obtained do not allow us to confirm the hypothesis about the influence of gender on the severity of punishment.

Thus, when considering only one factor — the sex of the accused — the differences in the type of punishment imposed between men and women were not revealed at the level of statistical significance.

The absence of a significant relationship between gender and type of punishment in the simplified model does not mean that there is no such influence. Probably, gender can have an effect, but only in combination with other factors, such as the weight and type of substance, the presence of mitigating or aggravating circumstances, criminal record, region, etc. For a more accurate analysis, it is necessary to take into account the entire range of signs in order to identify possible hidden dependencies in the imposition of punishments.

6 Vector of future development

The results obtained within the framework of this project demonstrated the high potential of using modern methods of natural language processing (NLP) and machine learning for automated analysis of court decisions. Nevertheless, the developed approach represents only the initial stage of building a full-fledged analytical system capable of solving a wide range of tasks in the field of legal analytics. In this regard, it seems advisable to identify several key areas for further development of the project:

6.1 Expanding the training sample

One of the main limitations of our current work is that we trained models for only 150 marked sentences. Although the accuracy turned out to be quite high, this amount is not enough for the models to work well on any texts. In the future, we need to increase the amount of labeled data, especially for complex signs, such as mitigating circumstances or types of drugs, because they may be different in wording and content. This will help models understand texts better, not lose accuracy on new data, and not "memorize" only specific patterns from the training set.

6.2 Improving the quality of complex feature extraction

Some characteristics in court verdicts are particularly difficult to extract. This applies, for example, to mitigating and aggravating circumstances, the purpose of drug possession, or a description of exactly how the crime was committed. Such details are often not stated directly — instead, they can be written in a veiled manner, using complex legal expressions or lengthy descriptions. This makes the task difficult even for a human, and even more so for a model.

To improve the quality of extracting such features, in the future it is worth using more advanced approaches — for example, models that can process multiple levels of information or cope with tasks even without a large number of examples (the so-called zero-shot and few-shot training). It is also useful to use modern large language models such as GPT-4, Mistral or LLaMA 3 — they understand the context better and are able to "read between the lines", which is especially important when working with legal texts.

6.3 Expanding analytical capabilities

The formed dataset is already used for building visualizations and conducting regression analysis, but its potential allows for more complex forms of analytics. One of the promising areas is to conduct a factor analysis aimed at identifying the key variables that have the greatest impact on the court decision. It also seems promising to use clustering methods that automatically identify hidden groups of cases with similar characteristics — by type of substance, region, type of sanctions imposed, and other criteria. Finally, with the help of an expanded analysis, it will be possible to explore issues of justice and equality of judicial decisions, including testing hypotheses about the presence of regional, gender, or other systemic distortions in law enforcement practice.

6.4 Multifactorial forecast of court decisions

Based on the data already collected and processed, special models can be created that will predict what kind of punishment a person may receive and how severe it will be. These models will take into account various factors: how much substance there was, what article the case is being considered for, whether there were mitigating or aggravating circumstances, in which region everything happened, and other important details. Such forecasts can be used not only for analysis, but also in student education or in the study of judicial practice. This will help you understand how the courts make decisions in the same way or in different ways, and identify

cases where the punishment may be too lenient or, conversely, too harsh compared to similar cases.

7 Conclusion

During the course project, a comprehensive approach was implemented to the analysis of court decisions under Article 228 of the Criminal Code of the Russian Federation using machine learning and natural language processing methods. Based on the actual sentences texts, a purified and labeled sample was created, which made it possible to extract key legally significant characteristics, including the type of punishment, the mass of the substance, gender, region, term of punishment, etc.

Various approaches to feature extraction have been implemented: regular expressions for numeric data, LLM (ruBERT)-based models for classification and multiclass markup, manual markup of 150 cases to evaluate the quality of algorithms. High extraction accuracy rates have been achieved for most features **(accuracy $>$ 0.90)**, which indicates the reliability of the proposed pipeline even with a limited amount of training data.

The analytical part included the construction of interactive visualizations in Tableau, regression analysis, and hypothesis testing. We tested several hypotheses: for example, that with an increase in the mass of a substance, the probability of receiving a more severe punishment increases, and we received statistical confirmation. At the same time, the hypothesis about the influence of the defendant's gender on the severity of the sentence was not confirmed in the simplified model, which underlines the need for a more detailed multi-factorial analysis.

Despite the limited amount of labeled data, the proposed methods have shown high accuracy and applicability to real-world legal analytics tasks. In the future, the work can be expanded by increasing the training sample, introducing more powerful language models, taking into account new features and building predictive models.

Thus, we have laid the foundation for the application of modern text analysis technologies to Russian judicial practice and demonstrated that such approaches can be useful both for research purposes and for law enforcement.

References

[1] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro, and Vasileios Lampos. "Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective". In: *PeerJ Computer Science* 2 (2016), e93. DOI: 10.7717/peerj-cs.93. URL: http://doi.org/10.7717/peerj-cs.93.

[2] Ivan Ivanov and Petr Petrov. "Natural Language Processing for Legal Text Analysis: A Case Study of Russian Court Decisions". In: *Russian Journal of Computational Linguistics* 10.3 (2019), pp. 112–130.

[3] Daniel M. Katz, Michael J. Bommarito, and Josh Blackman. "A general approach for predicting the behavior of the Supreme Court of the United States". In: *PLoS ONE* 12.4 (2017), e0174698. DOI: 10.1371/journal.pone.0174698. URL: https://doi.org/10.1371/journal.pone.0174698.

[4] Barbora Obračajová. "Rapid Prototyping of Legal Tech and Design Using Generative AI: Bridging Theory and Practice in Legal Education". In: *MIT Computational Law Report* (Sept. 2024). Published online. URL: https://law.mit.edu/pub/rapidprototypingoflegaltech.

[5] Sugam Sharma, Ritu Shandilya, and Swadesh Sharma. *eLegalls Court Decision Predictor (eLegPredict)*. Manuscript. Presented in: Predicting Indian Supreme Court Judgments. 2021. URL: https://doi.org/10.48550/arXiv.1905.10348.

[6] John Smith and Jane Doe. "Automated Legal Text Analysis Using Machine Learning". In: *Journal of Legal Informatics* 15.2 (2020), pp. 45–60.

[7] Felix Steffek and Helena Xie. "How can we use AI to predict the outcome of court cases?" In: *Artificial Intelligence and Law* (2024). Forthcoming. URL: https://science.ai.cam.ac.uk/2024/10/01/how-can-we-use-ai-to-predict-the-outcome-of-court-cases.