Analysis and Prediction of Legal Judgements in the Slovak Criminal Proceedings

Dávid Varga, Zoltán Szoplák, Stanislav Krajči, Pavol Sokol, and Peter Gurský

Institute of Computer Science
Faculty of Science, P.J.Šafárik University in Košice
Jesenná 5, 040 01 Košice, Slovakia
www.ics.science.upjs.sk
david.varga@student.upjs.sk, zoltan.szoplak@student.upjs.sk, stanislav.krajci@upjs.sk,
pavol.sokol@upjs.sk, peter.gursky@upjs.sk

Abstract: This paper uses machine learning to analyze criminal judgements in the Slovak republic to determine their adequacy and set a baseline for predicting their outcomes. First, we summarize past and recent advancements in predicting verdicts and other attributes of legal text written in different languages. We then demonstrate data preparation of all publicly available Slovak judgements, extraction of their verdicts and separation into main parts using a Slovak words inflexion dictionary called Tvaroslovník. Later we use this data to classify the judgements into acquittal or conviction using several known machine learning methods ranging from simple statistical methods such as SVM and random forests to deep learning networks based on convolution to recurrence and their combinations. We evaluate their efficiency, analyze and identify significant highly correlated terms with each result class, and offer a hypothesis as to why these terms are correlated with these results. We have found that a sequential input of word2vec embeddings combined with convolution-based deep learning methods produces the best results, achieving over 99% accuracy.

Keywords: judgement, reasoning, text analysis, Slovak, classification, verdict, machine learning

1 Introduction

Since 2016, the Ministry of Justice of the Slovak Republic has published more than 3 million publicly available court decisions online. These court decisions contain some structured data, e.g. name of the judge or court, but mostly free text. This free text contains the most relevant parts of court decisions: the final verdict and the reasoning behind the verdict. We aim to find a method to identify court decisions that are not sufficiently reasoned and provide such decisions to lawyers for a more detailed analysis.

In this paper, we examine several statistical and machine learning methods of text representation and classification, intending to correctly predict court decisions based on the reasoning alone.

After our model is trained, the reasoning and the verdict of the court decision will become inputs for this model. The model predicts the verdict from the input justification, comparing it with the true verdict received at the input. Subsequently, two situations can occur. If the predicted verdict is identical to the true verdict, we will take this court decision as sufficiently reasoned. If the predicted verdict differs from the true verdict, we will take such a decision as insufficiently reasoned. The model justifies its prediction by extracting the parts of the court's reasoning that most influenced the prediction of the verdict. This paper is based on the research stated in Sokol et al. [29], in which authors formulated their conclusions on the current state and developing trends in the use of digital evidence in judicial proceedings and usage of the *in dubio pro reo* principle in criminal proceedings.

To achieve a better understanding of how judgments are reasoned, this paper aims to:

- create a classification model which can predict the verdict of the judgement from its reasoning part;
- identify significant terms in the judgments' reasonings closely related to the results of judgements in the criminal proceedings (innocence or guilty).

This paper is organized into six sections. Section 2 focuses on the review of past and recent advancements in the classification of legal documents. Section 3 is devoted to data preprocessing and judgement extraction. Section 4 describes the different methods of text representation and the learning algorithms that will use them. The results produced by these algorithms and their subsequent analysis are presented in section 5, followed by the last section containing conclusions and future works.

2 Related works

2.1 A statistical approach

Predicting the results of court decisions from a statistical point of view was addressed by Kort [17] in 1957. He aimed to predict the cases concerning the right to counsel from The Supreme Court of the United States. He constructed a table with various facts of the cases paired with certain values. A composite value was calculated for each case by adding up all the facts' values. If the composite value of a particular case exceeded a certain threshold, then the defendant was wrongly denied the assignment of a lawyer. That way, he was able to predict successfully 12 of the 14 cases.

Later, Nagel (1960 [25] and 1963 [26]) applied the correlation analysis for court decisions. He predicted outcomes by calculating correlation coefficients for key variables, i.e. those which, according to the court, had the greatest influence on the determination of the judgment.

Mackaay and Robillard [22] applied the nearest neighbour rule method to predict judicial decisions, which was later verified by Keown [14] and compared with a linear model.

2.2 An approach based on artificial intelligence

In 2018, The European Commission for the Efficiency of Justice (CEPEJ) wrote the first European Ethical Charter on the use of Artificial Intelligence in judicial systems and their environment [7]. The charter summarized the basic principles that must be respected by artificial intelligence (AI). According to CEPEJ, AI can contribute to the efficiency of processing a large number of documents or resolving disputes. Still, it must be implemented responsibly, taking into account all human rights and personal data protection.

Currently, the most commonly used methods for predicting court decisions belong to machine learning, which is a part of artificial intelligence. Ruger et al. [28] and Katz et al. (2014 [12] and 2017 [13]) worked on a dataset from the United States Supreme Court called the Supreme Court Database (SCDB). They used methods such as the classification trees, the extremely randomized tree, LibLinear SVM and random forest.

Ashley et al. [3] were working on computer programs called SMILE and IBP that united case-based reasoning and information extraction from legal texts. By extracting information from previously decided cases, they attempted to predict the verdicts of new cases.

Aletras et al. [1] used machine learning to predict the rulings of the European Court of Human Rights (ECHR). Their data contained 584 judgments in English. They extracted main features from each decision using n-grams and trained the Support Vector Machines (SVM) classifier on these extracted n-grams. However, they did not remove the part of the decision in which the texts of the applicable laws were listed from the judgments. From such lists of laws, it was easier to predict the results of decisions. The success rate of classification was 79%.

In Medvedeva et al. [23], the authors decided to address this limitation while also dealing with the judicial decisions of the European Court of Human Rights. They removed the list of applicable laws from court decisions and used a larger number of decisions. The success of the classification deteriorated to 77%, using the same machine learning methods as Aletras et al. [1].

Another group of researchers, Chalkidis et al. [5], also predicted outcomes of decisions from ECHR using Bi-GRU with attention, hierarchical attention network and Label-Wise attention network. Attention scores provided indications of which part of the case affected the prediction the most.

Sulea et al. (2017) [31] decided to predict the verdicts of court decisions of the French Court of Cassation. They used a linear SVM classifier to train a bag of words instead of n-grams. They attempted to predict verdicts, the area of law and the length of court proceedings. Later that year, they [30] managed to increase the f1 score for each prediction using a system based on classifier ensembles.

Using the dataset from China Judgements Online (CJO), Luo et al. [21] attempted to predict the most frequent criminal charges and applied articles. The dataset already contained fact descriptions from which the authors extracted the applied articles by multiple SVM classifiers.

The previous dataset was also used by Hu et al. [10], but their task was a prediction of few-shot charges and a prediction of ten chosen attributes. They outperformed SVM, CNN, LSTM and the model created by Luo et al. [21] on few-shot charges by 50

In 2018, Xiao et al. [32] built a large dataset called CAIL 2018. It contains more than two million Chinese judicial decisions. The authors attempted to predict charges, applied articles and length of imprisonment only using baseline models such as SVM with TF-IDF, fastText [11] and CNN. To make predictions easier, they used only decisions with one defendant and decisions with frequent charges.

Using judgements from China Judgements Online, CAIL 2018 and Peking University Law Online, Zhong et al. [34] created a multi-task framework called TopJudge. It uses a directed acyclic graph for subtask dependencies and RNN for each subtask. Their subtasks were to predict applied articles, charges, fines and terms of penalty.

Long et al. [20] developed their own legal reading comprehension model named AutoJudge, which aims to model complex interactions among case materials and predicts the final verdict based on fact description, plaintiffs' pleas and law articles.

In 2020, Luz de Araujo et al. [2] created a new Brazilian dataset called VICTOR, which contained about 692,000 annotated legal documents. Legal experts annotated themes and a type of document (e. g. judgement and lower court decisions) about 6,800 documents which became a training dataset for further extraction. They used Naïve Bayes, SVM, BiLSTM and CNN for each type of classification, but the prediction of verdicts was not one of their tasks.

3 Dataset

The dataset presented in this work contained more than 3 million court decisions issued between 2016 and the end of

Table 1: Table of related works using the machine learning approach.

Literature	Datasets	Methods	Data representation
Ruger et al. (2004)	SCDB	classification trees	extracted variables
Katz et al. (2014)	SCDB	extremely randomized trees	extracted variables
Katz et al. (2017)	SCDB	random forest, LibLinear SVM, multilayer perceptron	extracted variables
Ashley et al. (2009)	custom database	SMILE + IBP	factor representation
Aletras et al. (2016)	ECHR	SVM	BoW, n-grams
Medvedeva et al. (2018)	ECHR	SVM	n-grams, TF-IDF
Chalkidis et al. (2019)	ECHR	BiGRU, HAN, LWAN, BERT, HIER-BERT	word embeddings
Sulea et al. (2017) [31]	The French Supreme Court	LibLinear SVM	BoW, n-grams
Sulea et al. (2017) [30]	The French Supreme Court	ensemble of multiple SVMs	BoW, n-grams
Luo et al. (2017)	CJO	custom method using SVM, softmax	sequence embeddings
Hu et al. (2018)	CJO	attentive attribute predictor, softmax	fact embeddings
Xiao et al. (2018)	CAIL 2018	SVM, fastText, CNN	skip-gram, TF-IDF
Zhong et al. (2018)	CJO, CAIL 2018, PKU Law Online	TopJudge	fact embeddings
Long et al. (2019)	CJO	AutoJudge	sentence embeddings
Araujo et al. (2020)	VICTOR	Naïve Bayes, SVM, BiLSTM, CNN, XGBoost	BoW, TF-IDF

2020. The court decisions covered all areas of legislation, such as civil, family, commercial and criminal law.

These court decisions are formatted as JSON objects, which contain attributes such as the type of the court, the name of the court, the name of the judge and the area of legislation. Each object has the document_fulltext attribute, which contains the anonymized court decision in its original version. During the preprocessing phase, we have been only working with this attribute and the area of legislation attribute.

There were several types of verdicts in these court decisions, such as the obligation to pay a sum of money, the acquittal of the defendant, the defendant's conviction, the rejection of the plaintiff's pursuit and many others. To simplify our work, we have decided to deal with criminal law containing a verdict of conviction and acquittal.

There were 226,500 court decisions concerning criminal law. We obtained these court decisions by searching for the value `"Trestné právo"` (criminal law) in the mentioned area of legislation attribute. From these decisions, it was necessary to extract the reasoning and the verdict, i.e. acquittal or conviction, which were used to train our models. After a more thorough filtration of these court decisions, explained in subsection 3 of this section, we ended up with 43,254 decisions with a conviction verdict and 3,139 decisions with an acquittal verdict.

3.1 Dividing court decisions into main parts

The part of the justification that is important for training the model was not present in the attributes of the original JSON files. Therefore, we have decided to split each judgment in its original form present in the document_fulltext attribute. We divided every judgment into these parts:

- details - contains semi-structured information about the court, the judge and the court decision. This information is the same as the values in mentioned attributes of JSON object;
- introduction - contains an introductory sentence in the judgment, the name of the court, the names of judges and defendants;
- statement - the section mentioning the verdict and the circumstances of the indictment;
- reasoning - the part in which the judgment is reasoned;
- judicial notice - instruction of the defendant, admissibility of the appeal and others.

The division of the judgments into mentioned parts was not problematic because their original texts were structured well.

Subsequently, we have replaced the original document_fulltext attribute of each JSON object with the newly created document_divided attribute, whose value was a JSON object with the attributes details, intro, statement, reasoning, and judicial notice.

3.2 Extraction of the verdict

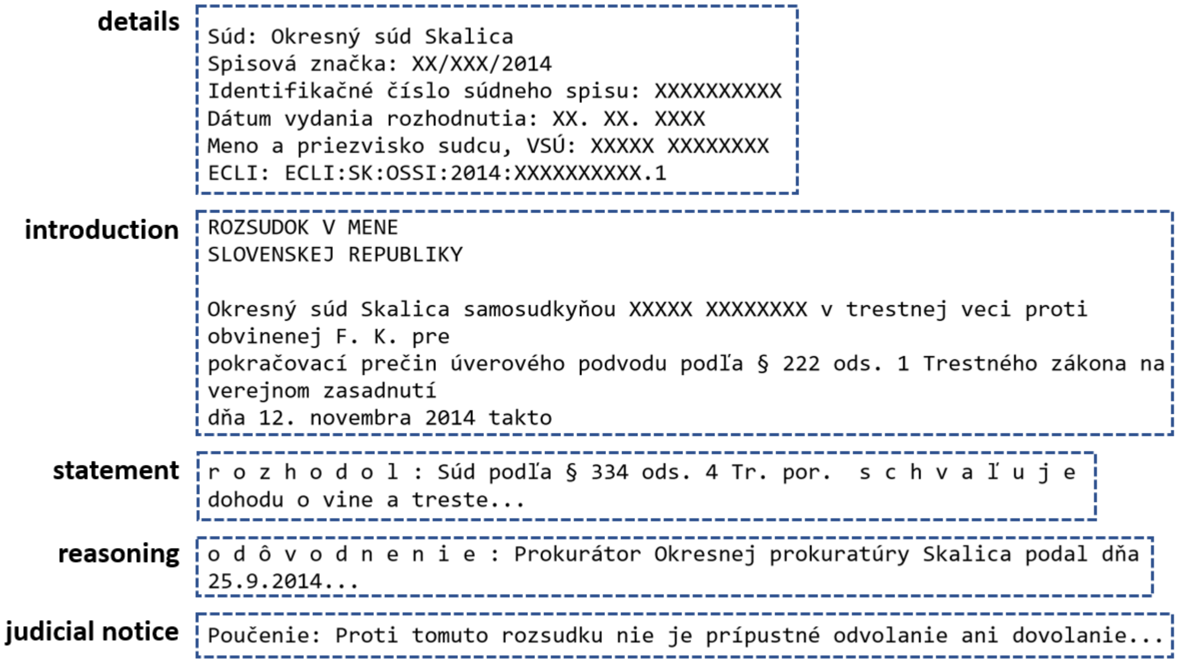
From observations, we have noticed that certain words are often spelt in a way that there is a space between each letter. The main verdicts were often written in this "spaced" style, e.g. the phrase `"j e   v i n n ý"` (is guilty). We decided to extract all words longer than two from the decision parts written in this style, and we wanted to find out how the conviction and the acquittal were formulated.

We created a finite state automaton to extract such words in one text pass. These words are then stored in the field of the newly created wide_words attribute.

The conviction always contained in its field wide_words the word starting with `"vinn-"`, i.e. the beginning part of the word `"vinný"` (guilty).

The acquittal always contained in the wide_words field a word starting with `"oslobod-"`, i.e. the beginning part of the word `"oslobodzuje"` (freed of charges).

Based on the occurrence and co-occurrence of these two terms, we divided the court decisions into four groups.



details

Súd: Okresný súd Skalica
Spisová značka: XX/XXX/2014
Identifikačné číslo súdneho spisu: XXXXXXXXXX
Dátum vydania rozhodnutia: XX. XX. XXXX
Meno a priezvisko sudcu, VSÚ: XXXXX XXXXXXXX
ECLI: ECLI:SK:OSSI:2014:XXXXXXXXXX.1

introduction

ROZSUDOK V MENE
SLOVENSKEJ REPUBLIKY

Okresný súd Skalica samosudkyňou XXXXX XXXXXXXX v trestnej veci proti
obvinenej F. K. pre
pokračovací prečin úverového podvodu podľa § 222 ods. 1 Trestného zákona na
verejnom zasadnutí
dňa 12. novembra 2014 takto

statement

r o z h o d o l : Súd podľa § 334 ods. 4 Tr. por. s c h v a ľ u j e
dohodu o vine a treste...

reasoning

o d ô v o d n e n i e : Prokurátor Okresnej prokuratúry Skalica podal dňa
25.9.2014...

judicial notice

Poučenie: Proti tomuto rozsudku nie je prípustné odvolanie ani dovolanie...

Figure 1: Segment of decision containing the verdict

The first group, named **none**, contained all court decisions in which neither the word beginning with "vinn-" nor the word beginning with "oslobod-" was mentioned. Such court decisions, for example, were requests for parole.

The second group, named **both**, contained court decisions which included in the court decision both a word beginning with "vinn-" and a word beginning with "oslobod-". Such court decisions often concerned several persons, several of whom were acquitted and others convicted.

The third group, named **guilty**, contained court decisions that contained words beginning with "vinn-" and did not contain a word beginning with "oslobod-". The fourth group named **innocent** contained words beginning with "oslobod-" and did not contain a word beginning with "vinn-". These two groups clearly define the verdict, and we used these two groups to train the model.

Due to the inconsistency of court decisions, it happened that a verdict was not written in "spaced" style but was written normally. For example, the verdict "j e v i n n ý" was written as "je vinný". We have also extracted these forms of verdicts by searching for words beginning with "vinn-" and "oslobod-".

3.3 Filtration of court decisions based on reasonings and verdicts

The first group of the court decisions that we excluded for the training set contained those that did not have the reasoning part. That is because, in certain cases, judges are not obliged to fill out the reasoning section. This group contained 130,289 decisions.

We also excluded those decisions that mentioned paragraph 172 article 2 of the Code of Criminal Procedure in their reasoning section. This article states that if both the prosecutor and the accused have waived their right to appeal or have made such a statement within three working days of the judgment, a simplified written judgment may be issued, not stating the reasons. This meant that even though the reasoning was present in the judgment, the reasoning itself stated that there is no justification stated in the judgment. We searched for the mentioning of this article using a regular expression and removed a further 15,953 judgements.

The last two groups removed from the training set were groups based on the type of verdict, specifically the **none** group, which contained 33,483 judgements and the **both** group which contained 382 judgements.

3.4 Further preprocessing

For each judgment, we have split the reasoning text into words and lemmatized them using a Slovak word form dictionary called *Tvaroslovník* described in [18]. We have also removed any non-alphabetic words and words shorter than three characters. We have used this text as the input and the verdict as the label. The data was split into a training and testing set, using two-thirds as training data. Due to the imbalance of target labels, we have downsampled the number of guilty verdicts in the training data to match the number of innocent examples.

4 Algorithms

4.1 Text representation

This section describes various representations of text and algorithms for predicting the outcome of court decisions. Most machine learning algorithms are incompatible with strings of characters as input data; thus, it is necessary to create numeric representations that preserve the syntactic and semantic relations between words.

A simple yet effective method of encoding is the **Tf-Idf** metric described in [27]. Tf-Idf (term frequency-inverse document frequency) is the combination of term frequency - the number of times a given term occurs within a document - and inverse document frequency - a metric that describes how unique or specific a given the word is to a document. Our vocabulary of terms contained not only individual words but also all bigrams and trigrams. This resulted in a large number of features even after excluding terms that occur less than five times total in the corpus. Therefore, we performed a χ^2 test to find the top 6500 terms that are most correlated with our target classes and used them as features calculating their Tf-Idf values for each document.

While effective, this kind of encoding does not tell us much about any spatiotemporal relations of the words themselves. Thus, we have opted to use vector embedding methods, namely **Word2Vec** and **Doc2Vec** which excel at encoding context for given words and documents. Word2Vec, described in [24] is a method for creating embeddings from each word by concatenating two prediction networks: CBOW, which tries to predict a word given the words surrounding it and Skip-Gram, trying to predict the surrounding words from the input word. We have trained a Word2Vec encoder with an embedding size of 300 on our dataset and used it in two distinct ways. We merely encoded each word of the padded judicial decisions for algorithms designed to work with sequential inputs. For algorithms that require encoding of the entire document, we calculated the element-wise mean, min and max values of all the word vectors of the decision. We concatenated them into an embedding with the size of 900. This simplistic method of pooling allows us to create a representation of a collection of words while still retaining semantic and syntactic information.

While the method above is somewhat effective, there is a more relevant method of creating embeddings from a sequence of words based on a similar principle, namely the Doc2Vec algorithm described in [19], a modification of the Word2Vec model to encode documents instead of words. Using this method, we have created an embedding of each judicial decision with a vector size of 500.

These representations can be used in conjunction with several machine learning algorithms to predict the verdict of judicial decisions.

4.2 Learning algorithms

There are several well-known if slightly outdated classifiers that have been used in NLP tasks that will serve as our baseline.

Logistic regression, as described in [16] is a method of classification that uses linear regression equations to produce discrete binary outputs.

A **Support vector machine**, described in [33] is an algorithm tasked with finding an optimal hyperplane that divides two or more classes with the greatest possible margin.

A **random forest**, described in [8] is a model that in itself is an ensemble of several decision trees.

These models can be used with representations that encode the reasoning as a singular input, meaning that the **Tf-Idf**, the concatenated **Word2Vec** and the **Doc2Vec** encodings can all be used.

In addition to these methods, we have decided to explore algorithms that use the sequence of words that make up the reasoning encoded by the Word2Vec method instead of taking in a singular input.

Convolutional Neural Networks or CNNs, described in [15] are based on the idea of using alternating layers of convolution - a sliding window function applied to a matrix - and pooling layers to subsample the input. While more well-known for their applications in computer vision, they can be applied to NLP tasks quite successfully due to their nature of capturing spatial dependencies and their ability to compose higher-level features from low-level features. We have used a single convolutional layer with 128 features and a kernel size of 5 with a maxpooling layer fed into a dense layer with ten neurons.

Recurrent Neural Networks or RNNs, on the other hand, have an internal state that can represent context information from an unspecified amount of past inputs. **Long Short Memory Networks** or LSTMs, described in [9] are able to deal with vanishing and exploding gradients better than traditional RNNs since they possess two gated units that open and close based on the relevance of the data, allowing it to better retain information over longer sequences. One shortcoming of conventional RNNs is that they are only able to make use of the previous context. Bidirectional RNNs are designed to process the data in both directions with two separate hidden layers, one processing the information going from the beginning forward in time and one from the end backwards. This approach allows us to have complete sequential information for each input about all points before and after. We use a single bidirectional block of LSTMs, each with 100 cells.

Some methods combine Recurrent Neural Networks with Convolutional Neural Networks in order to preserve both the spatial information retaining capabilities of convolutional networks and the temporal dependency capturing capabilities of recurrent networks.

The first is to create an ensemble model combining a convolutional network and a Bidirectional Gated Recurrent Unit described in [6]. The same input is presented to a CNN model with 100 features and a kernel size of 3 followed by a maxpool layer as well as a BiGRU model with a layer size of 64. The output of the two separately trained networks are concatenated into a single result.

Another, more indirect way of combining the attributes and strengths of RNNs and CNNs are **Temporal Convolutional Networks** or TCN networks, described in [4]. TCN use dilated causal convolution, meaning that outputs at time *t* is convolved only with elements from time *t* and earlier in the previous layer. This feature allows for parallel computation of convolutions rather than the sequential computation of RNNs and requires less memory than RNNs. As for the implementation, we will make use of 2 TCN blocks stacked with the kernel size of 3 and dilation factors of 1, 2, and 4, the first containing 128 filters the second 64 filters. The sequential output of the 2nd block is passed to 2 separate layers of pooling - max and average - the result of which is concatenated into a dense layer of 16 neurons then passed to the output.

In section 5, we describe the results of using these algorithms on the dataset described in chapter 3. Section 5.1 contains the evaluation of performance and subsequent comparison of these algorithms, whereas section 5.3 analyses what features and terms were used to make the predictions.

5 Results and discussion

5.1 Performance evaluation

We used the data described in section 3 and split it into three parts, using two for training and one for testing. We have implemented the methods described above and, after training, evaluated their performance using standard statistical metrics. These metrics consider the conviction samples as the `Positives` and the acquittal samples as `Negatives`. We have then organized these results into Table 2.

As we can see, regarding algorithms that use a singular representation(rows 1-9), the embedding models offer generally poorer performance, with the concatenated pooled Word2Vec being the least efficient since the algorithm is used in a way it is not designed to be used. Doc2Vec has better performance, especially when used in conjunction with Logistic Regression, where the relatively small number of features (500 as opposed to 900 and 6500) is less of a hindrance. However, the best results were achieved by using the Tf-Idf representation. We assume the reason for this is that the reasoning text has a somewhat formalized structure that uses certain standardized keywords and phrases from which basic information is more readily deductible than from a sequence of justifications presented within the reasoning.

This is somewhat further evidenced by the results obtained from methods reliant on the encoded sequence of words (rows 11-14). RNNs that are more heavily reliant on the sequential order of ideas have lower performance than CNNs, which have a property of location invariance thus are better suited to detect the presence of individual terms that are by large independent and highly correlated with the result class. The performance of such algorithms is quite high, achieving an accuracy of over 99%. We believe that this may be due to the relatively simple task of binary classification, combined with semi-structured data. We expect this to change as we try to predict more complex information from the dataset.

We can further observe from Table 2 that the precision for the prediction of conviction decisions is better than the recall metric for every single representation and model combination. Since precision is a metric that determines the percentage of predicted convictions to be actual convictions while recall tells us the percentage of actual convictions found by our algorithm, it stands to reason that a more significant number of convictions was classified as acquittal than the other way around.

Such bias may be the result of several possible causes. One of them is simply the consideration that there are suspicious cases within the dataset where the verdict should've been a conviction but ended up being acquittal. However, a more likely hypothesis is that many individual terms are highly correlated with the target classes and that many of them are, in actuality, more correlated with the conviction class of samples. So the decision process itself might try to detect values that are correlated more with conviction decisions, and upon their absence, it tends to classify acquittal. Unsure of the reason, we investigated what features contributed most to the prediction. Since embedding vectors are difficult to interpret, we used the feature selection method for the Tf-Idf representation using a bag of words and the χ^2 test. We calculated what percentage of documents from the training and testing corpus is the most relevant terms present for each target class. We organized these results into tables to determine which terms are used and how to make such decisions.

5.2 Definition of term categories

The terms (unigrams, bigrams and trigrams) can be divided into three categories according to their meaning and usage in a judicial decision:

- terms related to legal principles;
- terms used in legal arguments;
- other general legal terms, including terms describing the legal language.

The first group of terms is represented by terms related to the application of legal principles, resp. the exercise of rights under these principles. Judges often rely on legal principles to justify judicial decisions. An example is the principle of fair trial and the right to a fair trial.

Table 2: Table of classification results on the testing data. The rows represent the 14 different representation and algorithm combinations while the columns are the metrics we used to evaluate the performance of the given classifier.

Representation + Classifier	Accuracy	Precision	Recall	F1 score	ROC_AUC
word2vec + logistic regression	87.09	90.43	69.46	78.57	82.83
word2vec + svm	89.71	92.34	76.10	83.44	86.42
word2vec + random forest	94.87	99.52	85.35	91.89	92.57
doc2vec + logistic regression	97.83	98.10	95.34	96.70	97.21
doc2vec + svm	97.46	97.07	95.28	96.16	96.92
doc2vec + random forest	94.97	99.23	85.58	91.90	92.63
tf-idf +logitstic regression	95.64	98.69	88.18	93.14	93.80
tf-idf + svm	98.21	98.25	96.39	97.31	97.76
tf-idf + random forest	99.05	99.78	97.38	98.57	98.64
word2vec + CNN	99.24	99.78	97.89	98.83	98.89
word2vec + BiLSTM	98.72	99.20	96.83	98.00	98.23
word2vec + TCN	99.08	99.57	97.65	98.60	98.72
word2vec + Ensemble(CNN + BiGRU)	98.40	99.60	95.68	97.60	97.74

The second group consists of terms that are used in legal arguments. There are terms expressing usage and interrelationships of the evidence submitted in the criminal proceedings. Examples are general terms related to indication, such as to prove, proof. Another example is the use of evidence such as expert evidence, real evidence, documentary evidence.

The last group are general legal terms that do not fall into the groups mentioned above. These terms are part of the legal language and relate to legal institutes with a specific criminal offence (e.g. legal qualification, theft, breach of personal data protection), compensation or punishment. It also includes terms related to the procedure regulation of the court and law enforcement authorities (e.g. to accuse, hear, propose).

Certain legal principles are important for these proceedings, among which we can include the presumption of innocence of the defendant and the *in dubio pro reo* principle. This principle stipulates the obligation of the court to decide in favour of the defendant if there are doubts about his guilt that cannot be removed. It is this principle that creates a specific imbalance in thinking about guilt or innocence. The presumed result of judgement is innocence, and it is necessary to prove the defendant's guilt. It is a specific feature of the judgements in criminal proceedings, which is also reflected in the reasoning of the judgments. The judge needs to justify the guilt of the defendant and not his innocence.

5.3 Analysis of relevant terms

The second goal of this paper was to identify essential words or phrases associated with the decision on the merits in criminal proceedings. In other words, the aim was to determine the strength of the correlation between unigrams, bigrams and trigrams and the result in guilt or innocence.

In Table 3, we can see unigrams, bigrams and trigrams that have a significant relationship with judgments on the defendant's innocence with the chi-square value and the count of occurrences in judgements that point to the defendant's innocence or guilt. In contrast, Table 4 shows interesting unigrams, bigrams and trigrams, which are closely connected with judgements, the result of which is recognition of the defendant guilty. As we can see from these tables, specific terms correlate significantly more with the particular result of the judgement. Judges use in judgements' reasoning terms such as reason, unequivocally, female witness, situation etc. (Table 3) in the cases that result in acquittal of the defendant. On the other hand, expressions such as free choice, advise option, choice, voluntarily commit which, willingly etc. (Table 4) are important in the judgements condemning the defendant. The exciting finding is that groups of specific terms are closely connected with a specific type of verdict. The sets of terms prepared in this way can then be analyzed in terms of their mutual correlation or use as attributes for the classification of the judgements.

Within the used corpus of the judgements, we have focused on terms that are closely related to the evidence (evidence, prove, testimony, paper, etc.). The results show that these terms are strongly connected with judgements about the innocence of the defendant. Table 5 shows these terms with the chi-square value and the count of occurrences in judgements that point to the defendant's innocence or guilt.

These results suggest that for a judge to admit someone innocent, a much more detailed evidence-based argumentation must be used in the reasoning. At this point, it is necessary to return to the principle *in dubio pro reo*, which implies that the presumed result of judgment is innocence, and it is required to prove the defendant's guilt. It follows that the evidence and their representation in decision reasoning should be more closely linked to decisions with guilt verdict since guilt must be proved. However, here, we come to a disagreement between these claims and a dispute between the law in the book ("rules of the game"

Table 3: Table of terms relevant to judgments of innocence. The first column is a term in the Slovak language, the second column represents the translation of the term to English, in the third column Chi-square value is listed, and the last columns are the percentage of judgment on innocence, resp. guilt.

Term (Slovak)	Term (English)	Chi-square	Percentage_innocence	Percentage_guilt
obžalovať obžaloba	to charge the indictment	5295.07	39.6%	2.2%
svedkyňa	witness (female)	3026.24	37.6%	9.0%
pojednávanie	trial	2950.20	59.1%	22.4%
dôvod	reason	2756.17	48.3%	16.6%
jednoznačne	unequivocally	2653.22	33.0%	7.8%
prísť	come	2563.86	35.9%	9.8%
situácia	situation	2328.96	25.9%	5.2%
pamätať	to remember	2053.86	25.1%	5.6%
obdobie	period	1928.83	24.4%	5.8%
polícia	police	1920.33	29.8%	9.2%

Table 4: Table of terms relevant to judgments of guilt. The first column is a term in the Slovak language, the second column represents the translation of the term to English, in the third column Chi-square value is listed, and the last columns are the percentages of judgment on innocence, resp. guilt.

Term (Slovak)	Term (English)	Chi-square	Percentage_innocence	Percentage_guilt
slobodný voľba	free choice	9921.81	0.3%	34.4%
možnosť slobodný voľba	free choice option	9911.82	0.3%	33.8%
skrátiť vzdávať	to shorten give up	9530.09	0.0%	32%
súhlasiť návrh	to agree to a proposal	9452.09	0.0%	31.8%
radiť spôsob	advise option	9402.72	0.3%	32.4%
dobrovoľne spáchať	voluntarily commit	9261.44	0.3%	32.2%
voľba	choice	9241.86	0.9%	34.5%
dobrovoľne spáchať ktorý	voluntarily commit which	9157.98	0.2%	31.6%
dobrovoľne	willingly	5206.03	7.2%	37.1 %

for all cases) and law in action (judgment in the individual case). Based on the findings we have found, it appears that the judges do not presume the innocent of the defendant.

This specificity contained in the argumentation can then be seen in the algorithms that learn to recognize significant strings for two groups of decisions (guilty, innocent). This is evident from the precision and recall ratio as well as Table 4 and Table 3, where the higher χ^2 values and thus the features better suited for classification are correlated with judgements where the verdict was guilty. We have also calculated which of the top 300 terms occurs more in which class and have found that 223 of them had more occurrences in the guilty class, and only 77 had more in the innocent class. This supports the conclusion that we have arrived at after making observations from Table 2. At the same time, however, the conclusions of the paper [29], according to which more used evidence correlates with decisions on the innocence of the defendant, are confirmed. In the paper [29], authors focused only on the corpus of the judgements concerning digital evidence and IP addresses. In this paper, we use the extended corpus of the judgments, which covers various areas of criminal law.

6 Conclusion and future works

In this paper, we have shown how to split a judicial decision into its relevant parts and extract the verdict of the judgments. In addition, we have shown how to create a representation of the reasoning text using various text representation methods and combined them with several classification algorithms. We evaluated the performance of these models and found that methods that are more reliant on detecting specific terms than a stream of thoughts produce the most satisfactory results. Multiple models predict most cases with sufficient accuracy so that the outlying cases can be manually examined by a team of experts. Furthermore, it can be demonstrated that all representations and models are prone to classify conviction as acquittal more often than the other way around, which may be because our models tend to look for features present in convictions and interpret their absence as an acquittal.

As part of the analysis of significant terms, we have identified the groups of specific terms closely connected with a specific type of verdict (acquittal or conviction). Also, we have focused on the terms used in legal arguments (judgements' reasoning) in more detail. According to results, the *in dubio pro reo* principle in criminal proceedings affect judgement's reasonings and the subsequent

Table 5: Table of terms used in evidence-based argumentation. The first column is a term in the Slovak language, the second column represents the translation of the term to English, in the third column Chi-square value is listed, and the last columns are the percentages of judgment on innocence, resp. guilt.

Term (Slovak)	Term (English)	Chi-square value	Percentage_acquittal	Percentage_conviction
výpoveď	testimony	3652.45	52.1%	14.4%
preukázať	to prove	3071.27	52%	17.1%
dokázať	to prove	3064.15	26%	2.5%
dôkaz ktorý	evidence which	2929.94	28%	4%
výsluch	hearing	2839.05	42.2%	12.4%
listinný	documentary	2638.38	42.3%	13.3%
dôkaz	evidence	2625.49	55.3%	21.7%
listinný dôkaz	documentary evidence	2588.78	41.3%	13%
znalecký	expert	1998.48	28.6%	8%
dokazovanie	proving	1917.96	27.5%	7.9%

analysis of this legal text.

As an extension of this research, we plan to examine the cases where the labels and predictions differ and consult a lawyers team. Their task would be to determine for individual cases whether the failure is caused by the predictor, in which case we will research ways to improve our methods further. We will also replace all article references with the actual text of the articles to increase our predictive capability. We plan to make further predictions where in addition to determining the presence of guilt, we will also attempt to predict the severity of the sentence (e.g. jail time or fine amount). In case there are multiple defendants, we will try to determine the sentence for each of them.

References

[1] Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., Lampos, V.: Predicting judicial decisions of the european court of human rights: A natural language processing perspective. PeerJ Computer Science **2**, e93 (2016)

[2] Luz de Araujo, P.H., de Campos, T.E., Ataides Braz, F., Correia da Silva, N.: VICTOR: a dataset for Brazilian legal documents classification. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 1449–1458. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.181

[3] Ashley, K.D., Brüninghaus, S.: Automatically classifying case texts and predicting outcomes. Artificial Intelligence and Law **17**(2), 125–165 (2009)

[4] Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271 (2018)

[5] Chalkidis, I., Androutsopoulos, I., Aletras, N.: Neural legal judgment prediction in english. CoRR **abs/1906.02059** (2019), http://arxiv.org/abs/1906.02059

[6] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation (2014)

[7] European Commission for the Efficiency of Justice (CEPEJ): European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment (2018), https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c

[8] Ho, T.K.: Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition. vol. 1, pp. 278–282. IEEE (1995)

[9] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)

[10] Hu, Z., Li, X., Tu, C., Liu, Z., Sun, M.: Few-shot charge prediction with discriminative legal attributes. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 487–498 (2018)

[11] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models (2016)

[12] Katz, D.M., au2, M.J.B.I., Blackman, J.: Predicting the behavior of the supreme court of the united states: A general approach (2014)

[13] Katz, D.M., Bommarito, M.J., Blackman, J.: A general approach for predicting the behavior of the supreme court of the united states. PloS one **12**(4), e0174698 (2017)

[14] Keown, R.: Mathematical models for legal prediction. Computer/lj **2**, 829 (1980)

[15] Kim, Y.: Convolutional neural networks for sentence classification (2014)

[16] Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M., Klein, M.: Logistic regression. Springer (2002)

[17] Kort, F.: Predicting supreme court decisions mathematically: A quantitative analysis of the" right to counsel" cases. The American Political Science Review **51**(1), 1–12 (1957)

[18] Krajči, S., Novotný, R.: Tvaroslovník–databáza tvarov slov slovenského jazyka. In: Proceedings of international conference ITAT 2012. pp. 57–61. SAIA (2012)

[19] Lau, J.H., Baldwin, T.: An empirical evaluation of doc2vec with practical insights into document embedding generation. In: Proceedings of the 1st Workshop on Representation Learning for NLP. pp. 78–86. Association for Computational Linguistics, Berlin, Germany (Aug 2016).

https://doi.org/10.18653/v1/W16-1609, https://www.aclweb.org/anthology/W16-1609

[20] Long, S., Tu, C., Liu, Z., Sun, M.: Automatic judgment prediction via legal reading comprehension. In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) Chinese Computational Linguistics. pp. 558–572. Springer International Publishing, Cham (2019)

[21] Luo, B., Feng, Y., Xu, J., Zhang, X., Zhao, D.: Learning to predict charges for criminal cases with legal basis. CoRR **abs/1707.09168** (2017), http://arxiv.org/abs/1707.09168

[22] Mackaay, E., Robillard, P.: Predicting judicial decisions: The nearest neighbour rule. November, 1974 **41**, 302 (2020)

[23] Medvedeva, M., Vols, M., Wieling, M.: Using machine learning to predict decisions of the european court of human rights. Artificial Intelligence and Law **28**(2), 237–266 (2020)

[24] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)

[25] Nagel, S.: Using simple calculations to predict judicial decisions. American Behavioral Scientist **4**(4), 24–28 (1960)

[26] Nagel, S.S.: Applying correlation analysis to case prediction. Tex. L. Rev. **42**, 1006 (1963)

[27] Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 29–48. Citeseer (2003)

[28] Ruger, T.W., Kim, P.T., Martin, A.D., Quinn, K.M.: The supreme court forecasting project: Legal and political science approaches to predicting supreme court decisionmaking. Columbia Law Review pp. 1150–1210 (2004)

[29] Sokol, P., Rózenfeldová, L., Lučivjanská, K., Harašta, J.: Ip addresses in the context of digital evidence in the criminal and civil case law of the slovak republic. Forensic Science International: Digital Investigation **32**, 300918 (2020)

[30] Sulea, O., Zampieri, M., Malmasi, S., Vela, M., Dinu, L.P., van Genabith, J.: Exploring the use of text classification in the legal domain. CoRR **abs/1710.09306** (2017), http://arxiv.org/abs/1710.09306

[31] Sulea, O.M., Zampieri, M., Vela, M., Van Genabith, J.: Predicting the law area and decisions of french supreme court cases. arXiv preprint arXiv:1708.01681 (2017)

[32] Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H., Xu, J.: CAIL2018: A large-scale legal dataset for judgment prediction. CoRR **abs/1807.02478** (2018), http://arxiv.org/abs/1807.02478

[33] Zhang, Y.: Support vector machine classification algorithm and its application. In: International Conference on Information Computing and Applications. pp. 179–186. Springer (2012)

[34] Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., Sun, M.: Legal judgment prediction via topological learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3540–3549 (2018)