Legal Knowledge and Information Systems
S. Villata et al. (Eds.)

doi:10.3233/FAIA200850

Transformers for Classifying Fourth Amendment Elements and Factors Tests

Evan GRETOK [a] David LANGERMAN [a] and Wesley M. OLIVER[b]

[a] *Department of Electrical and Computer Engineering, Swanson School of Engineering, University of Pittsburgh, Pittsburgh, PA, USA*
[b] *Duquesne University School of Law, Pittsburgh, PA, USA*

Abstract. Determining if a court has applied a bright-line or totality-of-the-circumstances rule for Fourth Amendment cases demonstrates a difficult problem even for human lawyers and justices. Determining the type of test that governs an issue is essential to answering a legal question. Modern natural language processing (NLP) tools, such as transformers, demonstrate the capacity to extract relevant features from unlabelled text. This study demonstrates the effectiveness of the BERT, RoBERTa, and ALBERT transformer models to classify Fourth Amendment cases by bright-line or totality-of-the-circumstances rule. Two approaches are considered in which models are trained with either positive language extracted by a domain-expert or with full texts of cases. Transformers attain up to 92.31% accuracy on full texts, further demonstrating the capability of NLP techniques on domain-specific tasks even without handcrafted features.

Keywords. bright-line rule, totality-of-the-circumstances, fourth amendment, elements, factors, text classification, transformers

Introduction

To conduct legal reasoning, machines using artificial intelligence (AI) will have to identify the criteria the law uses to resolve an issue and extract evidence supporting those criteria. AI will also have to determine what the law does with those criteria to determine what sort of legal test is being used. In this paper, the authors show that, at least in the context of search and seizure law, it is possible for an automated system to examine a judicial opinion and identify the type of test used.

This research makes several contributions in determining the effectiveness of current natural language processing (NLP) systems to perform binary classification between bright-line and totality-of-the-circumstances rules, an important distinction in US criminal law. The authors perform transfer learning on several transformer models to extract meaning from the text. Models are fine-tuned on either key positive language extracted from cases by a domain expert or on the full text of the cases processed in a sliding-window approach. The accuracies of these models are compared with consideration to model size and complexity. The extraction of relevant language representation from full texts and successful classification of cases demonstrates the capability of current NLP systems to satisfy this need.

1. Background

This section outlines key concepts and related work. The legal background of bright-line and totality-of-the-circumstances rules is presented, key applications of AI to the law are revisited, and the technical aspects of transformer models are detailed.

1.1. Bright-Line and Totality-of-the-Circumstances Rules

Identifying the type of legal test a court is using is a fundamental question in resolving a legal issue. In a generic sense, legal tests are identified as either factors tests or elements tests. In an elements test, one seeking to obtain a legal remedy must satisfy all of the elements. In a factors test, a court will weigh the extent to which each of the factors is present, with the presence or absence of none of the factors being essential to resolve the issue either way [1].

Consider an elements test that requires a litigant to demonstrate *a*, *b*, and *c*. If there is no evidence on element *b*, then the litigant fails. It is much easier to resolve an issue governed by an elements test than one governed by a factors test. If a court is to consider factors *a*, *b*, and *c*, the absence of any support for factor *b* does not resolve the question. Similarly, evidence of *a*, *b*, and *c* would be sufficient to resolve an issue governed by an elements test, but not a factors test. Elements are either present or absent, factors must be weighed. Resolving a factors test is not beyond the capacity of a machine [2], though elements tests have proven easier for computers to analyze [3]. Regardless of the ease of the type of test, the machine must be able to deduce the type of legal test at issue.

Fourth Amendment cases were chosen because there are two types of clues in judicial opinions as to which sort of test the court is using. In other contexts, the choice between a test that considers factors or elements often is not driven by policy considerations and thus there is less likely to be professional commentary on the type of rule a court chooses to apply. In the Fourth Amendment context, a court's choice between a bright-line or totality-of-the-circumstances rule is very much a part of the discussion of academic commentators [4]. This constitutional provision governs searches and seizures. Bright-line tests provide clarity for police officers conducting investigations, but they also amount to judicially-created rule for the management of police. Totality tests defer to police departments for policy but provide little insight on what a court will find acceptable. In the case of New York v. Belton, the court concluded that if an officer had probable cause to arrest a motorist, the officer could search the entire car incident to arrest [5]. This is very simply an elements test with a single element. *Belton* did not ask whether the defendant was being arrested for a crime likely to yield evidence when the car is searched.

Belton demonstrates, however, that classifying a legal test is often a complicated question. There is a totality-of-the-circumstances test embedded within the Belton bright-line test. An officer must have probable cause to arrest a motorist. Probable cause is, of course, a totality-of-the-circumstances test. [6]. *Belton* is nevertheless regarded as a case that creates a bright-line rule. The Supreme Court was asked to consider whether a lawful arrest was sufficient to search the interior of a car, and the court determined that the right to search the interior of a car *always* accompanies the right to arrest a motorist.

There is an additional caveat complicating the classification of cases into one of these two groups. There are times when courts claim to be conducting a totality-of-the-circumstances test but are regarding a small set of facts that are likely to recur to clearly resolve the issue. Practically speaking, then, a commentator may label a test to be a bright-line rule while a court claims to be conducting a totality-of-the-circumstances analysis. Navarette v. California [7] is such an example. In *Navarette*, an anonymous informant reported improper driving and a police officer pulled the suspect over to ensure he was not drunk. Previously the Supreme Court had held in Florida v. J.L. that an anonymous tip that a person was possessing a gun was insufficient to detain the person identified [8]. Justice Thomas, speaking for the majority in *Navarette*, claimed to be applying a totality-of-the-circumstances test. Justice Scalia, who rarely disagreed with Justice Thomas, dissented, claiming that the majority had not been applying a totality-of-the-circumstances test at all, but rather creating a bright-line rule that anonymous tips of drunk driving were always sufficient to justify a stop [7]. For computational purposes, these cases would be identified as totality-of-the-circumstances cases even when, as a practical matter, a smaller number of commonly occurring factors prove to conclusively resolve the issue [9].

1.2. Basic Approaches to AI and Law

AI has become a mainstay of the legal profession. The ability for a computer model to process thousands of legal documents in minutes has reduced cost and fostered a new field of research. The field of AI and law has many facets, most of which lie in three areas. The first area is using AI to parse large corpora of text for relevant named entities [10], passages [11], or case law and statutes [12]. The second area is using AI to predict outcomes or behavior. This can take many forms such as predicting outcomes of court cases [13][14]. This area also includes the controversial topic of using AI to predict recidivism [15]. The final area is legal question answering and legal expert systems, where a large body of documents is used to train an AI to either directly answer questions or indicate logical paths of legal reasoning in search of fallacies or defenses [16].

Most approaches that filter or classify text rely on classical machine-learning (ML) methods that quantify some relationship of word or token frequency (i.e., bag of words representation) with a resultant label. This is done through the process of count vectorization, where a document is transformed into an embedding vector that uses unique words or n-grams as dimensions and their frequency of occurrence as the values for each dimension. These embedding vectors are then used as input to various ML models for prediction, such as a support vector machine (SVM), multi-layer perceptron (MLP), or decision tree (DT).

1.3. Deep Learning for Natural Language Processing

Recent NLP methods leverage an attention mechanism known as a transformer. Devlin describes the Google AI Language Lab's Bidirectional Encoder Representations from Transformers (BERT) model [17]. The effectiveness of this approach is its consideration of input sentences bidirectionally. This approach is borrowed from Vaswani in [18]. BERT requires no handcrafting of features and is able to extract meaningful representations directly from unlabeled text.

Liu presents a Robustly Optimized BERT Approach (RoBERTa). This research improves the training process of BERT and optimizes it via dynamic sentence masking.

Rather than training on recurrences of a sentence with a mask over a single word, the mask is moved to different words between training epochs. This allows the model to develop improved understanding of sentence structure and parts of language. The improvement was such that RoBERTa overtook BERT in capability for language understanding and question answering tasks [19].

Despite the capabilities of these models, each of them have on the order of a hundred million parameters and require many billions of operations to process texts. ALBERT, introduced in [20], attains similar accuracy at up to 18× lower parameter counts, as shown in Table 1, with a 1.7× reduction in training time. This was accomplished primarily by removing the dependency of the hidden layer and word embedding sizes and sharing parameters between layers. Larger variants of ALBERT, while still smaller than BERT, were able to attain a new state of the art on many of the same NLP benchmarks.

Table 1. Model Parameters and Layers [20][21]

Model	Variant	Parameters	Layers	Hidden Layer Size
BERT	Base Uncased	108M	12	768
BERT	Large Uncased	334M	24	768
RoBERTa	Base	125M	12	768
RoBERTa	Large	355M	24	768
ALBERT	Base v2	12M	12	768
ALBERT	Large v2	18M	24	1024

Transformer models have revolutionized deep learning for NLP. Their ability to capture relationships between distant segments of text helps them excel at complex tasks. Transformers have been used to expand the state of the art in benchmarks such as the Stanford Question Answering Dataset (SQuAD) [22], which asks an AI model to take an SAT-like test. Another challenging benchmark is reading comprehension, where an AI is asked to answer questions about a passage of text [23]. Transformer models consistently outperform the state of the art in these difficult tasks. In the law domain, transformers have been employed in recent work for judgement prediction [14], case law entailment [24], and legal news retrieval [25]. As the employed transformer models are limited to texts of up to 512 words, previous works consider hierarchical constructs of models for larger passages [14]. In many cases, this method is no longer required. A sliding-window approach to training existing transformer models on large text datasets can be enabled and customized with stride length parameters in the SimpleTransformers library [21].

2. Approach

This section details the experimental steps taken to make this research a reality. Key components include case preparation and model training.

2.1. Preparation of Cases

This experiment began with cases in the United States Supreme Court (SCOTUS) decided since 1946 [26]. WestLaw noted 880 Fourth Amendment cases decided by the US Supreme Court. A subset of these cases was identified in the literature as using or creating a “bright-line rule” or “totality-of-the-circumstances test.” Various law review articles described Fourth Amendment cases as fitting into one of the two models [27]. Unfortunately, The legal literature identified only a relatively small subset of the SCOTUS corpus as creating one of these two types of legal tests for interpreting the Fourth Amendment. Cases outside SCOTUS identified in the legal literature were therefore added to the dataset. The characterizations of the cases in the academic literature were accepted except when the literature took issue with the test courts claimed to be using [9]. In total, the dataset included 195 cases, 112 totality and 83 bright line.

The third author, a domain expert, then identified all case language deemed relevant to the court's analysis of the type of rule applied in or created by the case. An extensive inter-annotator agreement study was conducted in which each case was triply confirmed as bright line or totality by two independent legal citations and the opinion of the resident expert. Inter-annotation citations as well as positive and negative language from a small selection of cases can be referenced in Table 2. Finally, the full text of each case was extracted. Initial tests showed that roughly 200 cases was the minimum effective corpus size required for convergence in training. Hundreds of additional cases are available, but the time and expertise required to annotate data reduced labelling scope.

Table 2. Inter-Annotator Agreement with Key Positive and Negative Language

Case	Sources	Positive Language	Negative Language	Class
Ohio v. Robinette	[28] [29]	Voluntariness is a question of fact to be determined from all the circumstances.	...we have consistently eschewed bright-line rules...	Totality
Thornton v. United States	[30] [31]	Once an officer determines that there is probable cause to make an arrest, it is reasonable to allow officers to ensure their safety and to preserve evidence by searching the entire passenger compartment.	This determination would be inherently subjective and highly fact specific, and would require precisely the sort of ad hoc determinations on the part of officers in the field and reviewing courts that Belton sought to avoid.	Bright Line
Florida v. Royer	[32] [33]	All circumstances must be considered to determined whether someone is detained	We do not suggest that there is a litmus-paper test for distinguishing...	Totality
Alabama v. White	[34] [35]	We conclude that under the totality of the circumstances...	The Court there abandoned the "two-pronged test"	Totality
New York v. Belton	[4] [36]	A single, familiar standard is essential to guide police officers...	A custodial arrest of a suspect based on probable cause is a reasonable intrusion under the Fourth Amendment...	Bright Line

Preprocessing was conducted to prepare the text, enumerate classes, and split data into training folds. Texts were converted to lowercase as initial tests found improved accuracy without concern for proper nouns. This was also thought to minimize the effect of the relatively small dataset on the large transformer models. The effects of named-entity recognition were not considered in this research.

2.2. *Inter-Annotator Agreement*

Inter-annotator agreement was calculated to verify dataset validity. Cohen's kappa (κ) is the typical measure of agreement used to quantify how likely a dataset's distribution of agreement came about by chance rather than by true differences in the data. More on κ and how it is calculated can be found in the original paper by Cohen [37].

Table 3. Metrics Used to Calculate Cohen's Kappa for Inter-Annotator Agreement

Source A/Source B	Totality	Bright Line
Totality	106	0
Bright Line	12	77

$p_e = 0.509$, $p_O = 0.938$, $\kappa = 0.875$

As shown in Table 3, this dataset has a κ of ~0.87, implying near perfect inter-annotator agreement across all cited cases as to whether each represents totality-of-the-circumstances or bright-line rule.

2.3. Model Training

The SimpleTransformers library was employed for ease-of-access to pretrained BERT, RoBERTa, and ALBERT models [21]. The specific pretrained models bert-base-uncased, bert-large-uncased, roberta-base, roberta-large, albert-base-v2, and albert-large-v2 were fine-tuned in this study [38]. Transfer learning was conducted with three-fold cross validation to ensure effective fine-tuning on the full distribution of data. Three-fold was chosen as a 67% training set was sufficient for convergence but reduced the total training time. Training was conducted for twenty epochs for each model variant. Separate training rounds were considered for both domain-expert extracted positive language and full-text cases where the transformers were tasked with extracting language automatically. Training was accelerated on an NVIDIA TITAN RTX GPU using the Apex library. Final accuracy was fairly sensitive to initialization and dataset split, but this was expected due to the relatively small size of the dataset. For reference, the size of the combined full-text and positive language dataset used for transfer learning in this study is roughly ~4.3 MB of text, whereas a typical NLP dataset to train models of this size from scratch can range from tens of gigabytes to multiple terabytes [39] [40]. Accuracy referenced in the results section was computed as the mean of the F1-scores for each class.

The SimpleTransformers library provided a large number of parameters to tune the transfer-learning process. Do_lower_case was set to true as the dataset text was lowercased in preprocessing. FP16 precision was left enabled by default to increase training speed. To ensure full-text cases fit within the maximum 512 word model sequence length, the sliding_window parameter was set to true. The stride parameter was kept at its default 0.8. Default training and testing batch sizes of eight were used. Default values were used for the Adam optimizer epsilon, the learning rate, and the warmup ratio.

3. Results

This section details key results from this study, particularly the accuracy for each of the tested models. The authors' interpretation of these results follows.

3.1. Accuracy

Model accuracy results were determined for transformer base and large model variants on both domain-expert extracted positive language and full-text trials. These can be referenced in Table 4. Simple ML methods trained with the same positive-language and full-text datasets are included as a baseline for comparison.

Table 4. Model Accuracy (F1 Score)

Model	Variant	Positive Language	Full Text
Majority	Totality	62.00	62.00
SVM	sklearn	87.00	64.00
MLP	sklearn	88.00	76.00
Decision Tree	sklearn	69.00	56.00
BERT	Base Uncased	89.23	**92.31**
RoBERTa	Base	87.18	90.26
ALBERT	Base v2	87.69	91.79
BERT	Large Uncased	86.15	91.79
RoBERTa	Large	**90.26**	78.97
ALBERT	Large v2	81.03	57.44

These results demonstrate that transformer models can perform high-accuracy classification of cases by both positive language and full text. For previous research, the process of handcrafting the dataset was often an essential step. These results show that, while domain-expert extracted positive language may yield good accuracy, transformers enable full-texts provide an even better result. Simple ML methods simply cannot compete on full texts. The transformer models are able to extract adequate feature representations from the text on their own without human intervention. This showcases the value of using modern NLP techniques for this and similar problems in the legal domain.

The smaller transformer models performed very consistently at around 88% accuracy for positive language and 92% for full texts. In many cases, the reduced parameter counts, training times, and inference costs of the base models may make them a more attractive solution. The large models are considerably less consistent and often perform worse. For this study, one factor may be the relatively small dataset. The large model variants contain many more parameters to tune. Without a large dataset exercising these parameters during the transfer-learning process, the model becomes data starved and is not as effective. Overfitting may also take place if large models are allowed to train for too long, though further investigation is necessary to see if a double-descent phenomenon is presenting itself in this case [41].

In comparing transformer model types, differences again arise when considering the larger variants, especially for full texts. The reduction in performance for the large RoBERTa model is perhaps best attributed to the lack of cased data. While BERT has an uncased model, which was used in this study, RoBERTa does not. Some components of the pretrained model that applied specifically to cased data may never have been activated or utilized here, reducing the model's overall effectiveness. ALBERT may suffer from the opposite effect. As a smaller, more optimized model, it may simply not have the parameter space required to correctly filter the barrage of features from the full-text cases. Embedding is the key component for this task, and the methods used to reduce size and compute expense for ALBERT have reduced its capability here. Even larger ALBERT variants, such as ALBERT-XXLarge, may be able to overcome the simpler embedding limitations, but were unfortunately beyond the scope of this research.

3.2. Analysis

Proper transfer learning of transformers is about much more than just the quantity of samples. The quality of data fed to the model for training is a considerable factor. It is a double-edged sword. If good text is supplied and an accurate representation is extracted quickly, additional training epochs may result in degradation of the model and loss of accuracy. Conversely, if the text supplied is poor in representation or limited in scope, the model may struggle with extracting a representation at all, resulting in alarmingly poor metrics for the same training period and parameters.

While this study attained high peak accuracy and saw convergence of nearly all model types, the accuracy between training runs varied considerably in some cases. This is likely due to the variance in the amount and quality of language between different cases. Three-fold cross validation reduced this variance by ensuring that the full dataset could be used in each training round. The results of training are understandably more biased by the presence or absence of proper language resources in training than the number of cases in the sample for each class. Proper balance of the training dataset for abstract text classification tasks is critical, especially when training large models with a relatively small dataset. Even slight bias may lead to overfitting and a decrease in testing accuracy for the underrepresented class. This was experienced in many of the sub-par accuracy transformer training rounds. With such a high number of interrelated parameters to tune, biases in the subset of the small dataset selected for training quickly become apparent. This caused a small number of training rounds to fail to converge. The authors emphasize that these results demonstrate a proof of concept. For this approach to be employed in a production tool, a larger dataset would have to be prepared.

4. Conclusions

This research demonstrates that an automated system can be taught to identify whether a court is using or creating an elements or factors test. Reproducing this experiment in other substantive areas will of course require some modification to these methods. Outside the Fourth Amendment context, there is less discussion, in judicial opinions or the academic literature, about the choice between a multi-factor totality-of-the-circumstance test and a clearer test that turns on whether a small set of criteria are fully satisfied or not. Fourth Amendment cases will therefore include more language than cases in many other contexts which automated systems may use to assess the type of cases used. Academic journals less frequently discuss the type of rule chosen in other contexts, providing less readily available annotation. Nevertheless, the very high degrees of accuracy obtained in the Fourth Amendment context suggests that transformer models are capable of differentiating the type of legal test used in a legal opinion, an effective first step.

4.1. Key Accomplishments

Correct classification of bright-line and totality-of-the-circumstances cases is achievable with current transformer-based NLP methods. Fine-tuned BERT, RoBERTa, and ALBERT models were successfully employed for binary classification of full texts in this study. Deep-learning transformers attained accuracies of up to 90.26% on positive language and 92.31% on full texts. This research demonstrates the scalability of transformers to longer lengths of text via a sliding-window approach. The results show that, while positive language is sufficient, transformer models are now capable of extracting their own effective feature representations from supplied text to perform at even higher accuracy. The process of fine-tuning a pre-trained transformer for text classification is shown as one attainable and accessible method for assistive AI in a variety of legal domains.

4.2. Future Work

Similar methods may be applied to a larger dataset. Continuing to grow the corpus should increase model accuracy. Further exploration to compare and contrast effectiveness of different transformer models and variants is merited. A study of the effect of cased and uncased language could also provide insight. A more thorough assessment of different learning rates, batch sizes, and other parameters could be effective, but was not within the scope of this study.

Acknowledgements

The authors would like to acknowledge the contribution of Rui Zheng to case aggregation and labelling tasks. This research began in satisfaction of a course project for the University of Pittsburgh School of Law's LAW 5719 Applied Legal Data Analytics and AI course taught by Dr. Kevin Ashley and Dr. Jaromir Savelka. The authors are very thankful for their insight, proofing, and expertise in the completion of this paper.

References

[1] Smith M. Advanced legal writing : theories and strategies in persuasive writing. New York: Aspen Law & Business; 2002.

[2] Ashley KD. Artificial intelligence and legal analytics: New tools for law practice in the digital age. Cambridge University Press; 2017.

[3] Gardner AvdL. An Artificial Intelligence Approach to Legal Reasoning. Cambridge, MA, USA: MIT Press; 1987.

[4] Alschuler A. Bright Line Fever and the Fourth Amendment. University of Pittsburgh Law Review. 1984 1;45.

[5] New York v. Belton, 453 US 454 - Supreme Court 1981;.

[6] Dery III GM, Dery GM. Issue 3 2000 III, Improbable Cause: The Court's Purposeful Evasion of a Traditional Fourth Amendment Protection in Wyoming v. Houghton, 50 Case W. Res; 2000.

[7] Navarette v. California, 572 U.S. 393, 2014.;.

[8] Florida v. JL, 529 US 266 - Supreme Court 2000;.

[9] Nelson DM. Illinois v. Wardlow: A Single Factor Totality. Utah Law Review. 2001;2001.

[10] Cardellino C, Alemany LA, Teruel M, Villata S. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In: Proceedings of the International Conference on Artificial Intelligence and Law. New York, New York, USA: Association for Computing Machinery; 2017. p. 9–18.

[11] Šsavelka J, Ashley KD. Segmenting U.S. court decisions into functional and issue specific parts. In: Frontiers in Artificial Intelligence and Applications. vol. 313. IOS Press; 2018. p. 111–120.

[12] Koniaris M, Anagnostopoulos I, Vassiliou Y. Network analysis in the legal domain: A complex model for European Union legal sources. Journal of Complex Networks. 2018 4;6(2):243–268.

[13] 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. 2016 10;2016(10):e93.

[14] Chalkidis I, Androutsopoulos I, Aletras N. Neural Legal Judgment Prediction in English. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. 2019 6:4317–4323.

[15] Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias; 2016.

[16] Ashley KD. Case-Based Reasoning and its Implications for Legal Expert Systems; 1992.

[17] Devlin J, Chang MW, Lee K, Google KT, Language AI. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding;.

[18] Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention Is All You Need;.

[19] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach; 2019.

[20] Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In: Proceedings of the International Conference on Learning Representations; 2020. .

[21] ThilinaRajapakse/simpletransformers: Transformers for Classification, NER, QA, Language Modelling, Language Generation, T5, Multi-Modal, and Conversational AI;.

[22] Rajpurkar P, Jia R, Liang P. Know what you don't know: Unanswerable questions for SQuAD. In: ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). vol. 2. Association for Computational Linguistics (ACL); 2018. p. 784–789.

[23] Lai G, Xie Q, Liu H, Yang Y, Hovy E. RACE: Large-scale ReAding comprehension dataset from examinations. In: EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings. Association for Computational Linguistics (ACL); 2017. p. 785–794.

[24] Rabelo J, Kim MY, Goebel R. Combining similarity and transformer methods for case law entailment. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law; 2019. p. 290–296.

[25] Sanchez L, He J, Manotumruksa J, Albakour D, Martinez M, Lipani A. Easing Legal News Monitoring with Learning to Rank and BERT. In: Proceedings of the European Conference on Information Retrieval; 2020. p. 336–343.

[26] The Supreme Court Database;.

[27] Sommers CD. Presumed Drunk until Proven Sober: The Dangers and Implications of Anonymous Tips Following Navarette V. California. South Dakota Law Review. 2015 6;60(2):327.

[28] Katz LR. Baldwin's Ohio Handbook Series: Ohio Arrest, Search and Seizure. 2020;20(2).

[29] Mendelson A. The Fourth Amendment and Traffic Stops: Bright Line Rules in Conjunction with the Totality of the Circumstances Test. J Crim L and Criminology. 1988;88:930.

[30] Glandon KR. Bright Lines on the Road: The Fourth Amendment, The Automatic Companion Rule, the "Automatic Container" Rule, and a New Rule for Drug- or Firearm-Related Traffic Companion Searches Incident to Lawful Arrest. Am Crim L Rev. 2009;46:1267.

[31] Butterfield EJ. Bright Line Breaking Point: Embracing Justice Scalia's Call for the Supreme Court to Abandon an Unreasonable Approach to Fourth Amendment Search and Seizure Law. Tul L Rev. 2007;82:77.

[32] Saleem O. The Age of Unreason: The Impact of Reasonableness, Increased Police Force, and Colorblindness on Terry "Stop and Frisk". Okla L Rev. 1997;50:451.

[33] Urbonya KR. Rhetorically Reasonable Police Practices: Viewing the Supreme Court's Multiple Discourse Paths. Am Crim L Rev. 2003;40:1387.

[34] Bryk JK. Anonymous Tips to Law Enforcement and the Fourth Amendment: Arguments for Adopting an Imminent Danger Exception and Retaining the Totality of the Circumstances Test. Geo Mason U Civ Rts L J. 2003;13:277.

[35] Krippandorf E. Florida v. J.L.: To Frisk or Not to Frisk: The Supreme Court Sheds Light on a Use of Anonymous Tipsters as a Predicate for Reasonable Suspicion. New Eng J on Crim and Civ Confinement. 2002;28:161.

[36] Dripps DA. Responding to the Challenges of Contextual Change and Legal Dynamism in Interpreting the Fourth Amendment. Miss L J. 2012;81:1085.

[37] Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960 4;20(1):37–46.

[38] huggingface. Pretrained Models. 2020.

[39] Klimt B, Yang Y. The enron corpus: A new dataset for email classification research. In: Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science). vol. 3201. Springer Verlag; 2004. p. 217–226.

[40] Google Books Ngrams - AWS Public Data Set;.

[41] Nakkiran P, Kaplun G, Bansal Y, Yang T, Barak B, Openai IS. Deep Double Descent: Where Bigger Models and More Data Hurt. In: ICLR 2020; 2020. .