# Evaluating Accuracy of Sentence Classification in Legal Documents using Transformers

Samatva N Kasat, Chirec International School, Hyderabad, India
samkas125@gmail.com
Research Report under the guidance of
Dr. Krishnan Pillaipakkamnatt, Professor, Department of Computer Science,
Hofstra University, New York, USA
krishnan.pillaipakkamnatt@hofstra.edu

**ABSTRACT**

This paper focuses on evaluating the accuracy of sentence classification transformer models to classify sentences from legal documents into different rhetorical roles. Rhetorical role classification can be applied to mine patterns of reasoning from large legal documents. Use cases include semantic viewers, semantic search, decision summarizers, argument recommenders, and reasoning monitors. Through this application, legal services can be made more efficient and thus a common person's access to justice can be made affordable. In this paper, we compare the accuracy of sentence classification using BERT-based transformers with previous approaches like rule-based methods and standard machine learning methods.

## 1. Introduction

The use of Artificial Intelligence (AI) in the legal domain has been explored for several decades. Over the past few years, however, there has been increased debate regarding the degree of usage of Machine Learning (ML) tools in the legal field. While most researchers dismiss the possibility of AI completely replacing human lawyers, it is widely agreed that AI will play a significant role in law as a tool to assist lawyers. With the advent of large language models, the debate about the extent of AI usage in the legal domain has gained significant attention. It is certain that, with time, AI will become a key tool within the domain of law.

In fact, AI is already in use in the legal system. Chatbots, for example, are used to answer questions from the public about standard operating procedures, manuals, and other existing information resources with a high degree of accuracy. In India, the Supreme Court Portal for Assistance in Court Efficiency facilitates the judge's research by processing relevant facts and relevant laws. AI is also used to translate legal documents [5]. There are several non-public initiatives in Germany to develop solutions for analyzing existing decisions and inferring what a decision could look like in the case at hand [3].

Currently, law firms hire an army of para-legals to assist lawyers in mining arguments from previous arguments and judgements. Automating this process can enable them to mine a lot more data at a much lower cost. This would provide lower cost law firms with high quality resources, giving them the ability to compete effectively against larger and more established law firms. In certain areas in the legal industry, particularly legal research, contract review, e-discovery, and investigations, ML products have already been developed and marketed [11].

However, improving accuracy is critical to having faith in these tools. In fact, being able to measure the accuracy of these tools, something which is not possible for human-conducted legal research, gives researchers and lawyers greater control, and leads to higher scrutiny of the technology. Ultimately, this allows for the development of highly accurate and reliable models.

Improving the accuracy of AI-based legal approaches requires large amounts of accurately labelled data. This problem is exacerbated for unprivileged sections of society, which tend to have lower amounts of available data for research and training. This is because underprivileged sections of society have traditionally utilized the

legal system less than other groups, resulting in fewer available legal documents.

In this paper, we focus on improving the accuracy of legal sentence classification by using latest transformer models. We aim to provide additional data points for further research. We make use of a publicly available annotated dataset at https://github.com/LLTLab/VetClaims-JSON [2].

After we discuss prior work in Section 2, we will provide an overview of the datasets and the ML system in Section 3 and show a comparison of the results of our work with prior work on sentence classification in Section 4. Finally, we provide a conclusion and possible future work in this area in Section 5.

## 2. Related Work

Text classification is a machine learning problem that deals with categorizing any given input text into a set of predefined categories. A vast majority of textual data available is unstructured and it is time consuming to extract meaningful information out of it. Attempts to automate text classification have been done using various approaches using rule-based system and various machine learning based systems like Naïve Bayes (NB), Logical Regression (LR), and Support Vector Matrix (SVM) [8, 9, 10]

This paper addresses text classification of legal documents, and it references work being done in [1] which presents quantitative results of various approaches of automated argument mining from adjudicatory legal decisions. Such argument mining would automatically extract the evidence assessment and fact-finding reasoning found in adjudicatory decisions, for the purpose of identifying successful and unsuccessful units of evidentiary argument. Previous work has also been done in classifying roles of sentences for argument mining using various machine learning approaches like NB classifier, LR classifier and SVM for the Dutch Laws [7]. Each has yielded varying degrees of accuracy thus providing insights into the data. Researchers have tried to exploit the well-defined structure of the legal sentences to explore whether a small sample of data can be used and apply a rule-based approach. This has yielded some encouraging results. However, further improvements in accuracy, along with error reduction are required to achieve the degree of reliability needed for real-life use cases. Furthermore, using ML-based approaches may facilitate the creation of a more generalized model, which may yield promising results in other types of legal cases as well.

## 3. Datasets and System Overview

### 3.1 Datasets

This paper makes use of Board of Veterans' Appeals ("BVA") of the U.S. Department of Veterans Affairs PTSD dataset available at https://github.com/LLTLab/VetClaims-JSON [2]. As part of this data set, sentences are classified and labelled into 5 categories. This is available in a JSON file. The five rhetorical roles for sentences used in these JSON decision models are:

*FindingSentence (F)*: A finding sentence is an official determination by a jury or judge of a fact or conclusion based on evidence presented in court.
*EvidenceSentence (E)*: An evidence sentence describes the testimony of an evidence bearer, describes documents that may be used as evidence, or details any other type of evidence.
*ReasoningSentence (R)*: A reasoning sentence reports the judge's or jury's reasoning, or evaluation of the evidence presented in order to reach a finding.
*LegalRuleSentence (L)*: A legal rule sentence states or explains one or more legal rules. It only contains a "matter of law" and may not contain a "matter of fact".
*CitationSentence (C)*: A citation sentence states a reference to a legal authoritative document, source, or precedent. It generally includes standard notation to describe the cited source.
*OtherSentence (S)*: Other sentences refer to sentences which do not classify as any of the aforementioned categories.

Table 1: Distribution of various sentence types in the BVA PTSD dataset

| Rhetorical type | Frequency |
|---|---|
| Finding Sentence | 490 |
| Reasoning Sentence | 710 |
| Legal-Rule Sentence | 938 |
| Citation Sentence | 1,118 |
| Other Sentences | 478 |
| Total | 6153 |

In our setup, out of total 6153 sentences, we used 5000 datasets for training sentences and used the remaining sentences to test the accuracy of the model. Given the relatively small dataset, we decided to use a greater number of datapoints for training.

## 3.2 ML System Overview

Transformers are neural networks that learn context and understanding through sequentially analyzing data [6]. The transformer models use a set of mathematical techniques, generally known as attention or self-attention [6]. These techniques help identify how distant data elements influence and depend on one another. Through this, a model can "understand" long sequences of data with accuracy. Transformers draw inspiration from the encoder-decoder architecture. However, they do not perform data processing in sequential order, thus allowing for greater parallelization and faster training. Transformers' attention mechanism is an essential component, which indicates the importance of encoding a given token in the context of other tokens in the input.

Given the large computation power required for training, there is a repository of pre-trained transformer models available at HuggingFace. These pre-trained models can then be further "fine-tuned" on a specific set of data at a relatively low cost. This enables broader adoption of transformers. In this research, we use HuggingFace's bert-large-cased transformer model [4] to do sentence classification.

There are newer and more powerful transformer models also available. But we chose this model to find a balance between computational restrictions due to the machine used, and the accuracy of classification. Also, using newer and larger transformer models could lead to over-fitting, although this could be a topic of future research. BERT is a transformer model pretrained on a large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labelling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. This model has 24 layers, 1024 hidden dimensions, 16 attention heads, and 336M parameters.

We further fine-tuned the model with the data obtained from [2]. The fine-tuning was done on a machine with NVIDIA GTX 1650 Ti GPU with 4 GB RAM. This was a relatively less powerful machine, however, given our small dataset, it proved to be sufficient. The fine-tuning took 5 epochs to converge on the accuracy and took approximately 24 hours to complete the fine-tuning.

In order to gain more data, we also tried text augmentation. However, it did not help improve accuracy. This is likely because the legal sentences follow a very well-defined structure. Thus, any generic text augmentation approach can lead to introduction of new sentence structure and vocabulary not used in judicial judgements. Hence, the additional variety does not help improve the accuracy of classification. This dependence on large quantities of training data is one aspect in which ML-based approaches may not be as effective as rule-based scripts.

## 4. Results

In this section, we present our results. We note that our work shows that transformers provide an improved accuracy of 88% as compared to the previously achieved highest value of 85.7% using other techniques. This corresponds to a 16% reduction in error, a noticeable improvement from previous work. Below is a table of accuracy results of our work with other approaches. We also record the confusion matrix with both raw numbers and in terms of percentage.

The fine-tuned model has been uploaded and can be accessed through HuggingFace repository at https://huggingface.co/samkas125/bert-large-legal-sentence-classification.

Table 2: Comparison with other approaches

| Algorithm/ Metrics | Multi-class accuracy | Multi-class false positive |
|---|---|---|
| NB | 81.7 % | 1.5 % |
| LR | 85.7 % | 1.6 % |
| SVM | 85.7 % | 1.6 % |
| Bert-large-cased | 88.0 % | |

Table 3: Bert-large-cased Confusion Matrix – Raw Numbers

| | S | F | E | R | L | C |
|---|---|---|---|---|---|---|
| **S** | 85 | 1 | 3 | 8 | 3 | 0 |
| **F** | 2 | 82 | 5 | 12 | 0 | 0 |
| **E** | 2 | 6 | 444 | 16 | 3 | 0 |
| **R** | 4 | 10 | 30 | 77 | 11 | 1 |
| **L** | 5 | 1 | 0 | 11 | 146 | 1 |
| **C** | 0 | 0 | 0 | 0 | 1 | 183 |

Table 4: Bert-large-cased Confusion Matrix – Percentage

| | S | F | E | R | L | C |
|---|---|---|---|---|---|---|
| **S** | 0.85 | 0.01 | 0.03 | 0.08 | 0.03 | 0 |
| **F** | 0.02 | 0.81 | 0.05 | 0.12 | 0 | 0 |
| **E** | 0.0042 | 0.013 | 0.94 | 0.034 | 0.0064 | 0 |
| **R** | 0.03 | 0.075 | 0.23 | 0.58 | 0.083 | 0.0075 |
| **L** | 0.03 | 0.0061 | 0 | 0.067 | 0.89 | 0.0061 |
| **C** | 0 | 0 | 0 | 0 | 0.0054 | 0.99 |

Based on the confusion matrix, we calculate additional metrics.

Table 5: Transformer – Measure of Accuracy

| | TP | FN | TN | FP | FPR | TNR |
|---|---|---|---|---|---|---|
| **S** | 0.85 | 1 | 4.9135 | 0.0842 | 0.0168 | 0.9831 |
| **F** | 0.81 | 0.19 | 4.8936 | 0.1041 | 0.0208 | 0.9791 |
| **E** | 0.94 | 0.94 | 4.6901 | 0.31 | 0.0619 | 0.9380 |
| **R** | 0.58 | 1.0055 | 4.6912 | 0.301 | 0.0602 | 0.9397 |
| **L** | 0.89 | 0.9992 | 4.8737 | 0.1248 | 0.0249 | 0.9750 |
| **C** | 0.99 | 0.9954 | 4.9887 | 0.0136 | 0.0027 | 0.9972 |

TP: True Positive.
FN: False Negative
TN: True Negative
FP: False Positive
FPR : False Positive Rate = FP / (TN + FP)
TNR : True Negative Rage = 1 - FPR

Charts below show a comparison of various precision, recall and F1-score between bert-large-cased (Transformer), Logical Regression (LR) and Support Vector Machine (SVM) based classification. We note that bert-large-cased gets better or comparable precision in majority of the classes except reasoning. Similarly, it also achieves better or comparable recall value in all classes except legal rules.

Chart 1: Comparison of Precision between various AI techniques

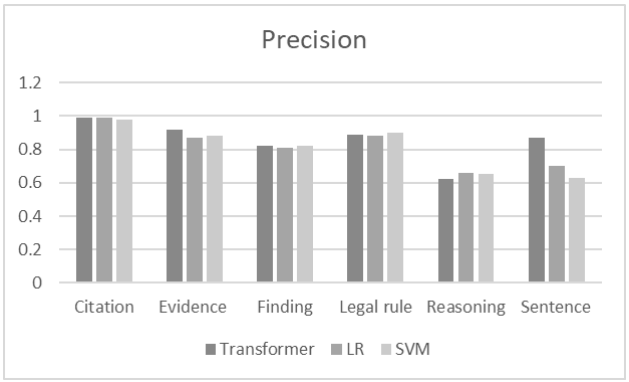



Chart 2: Comparison of Recall values between various AI techniques

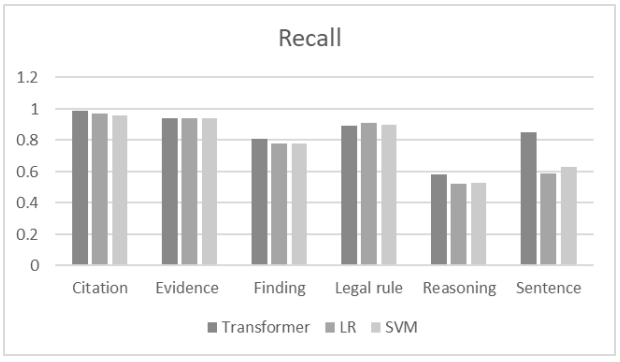
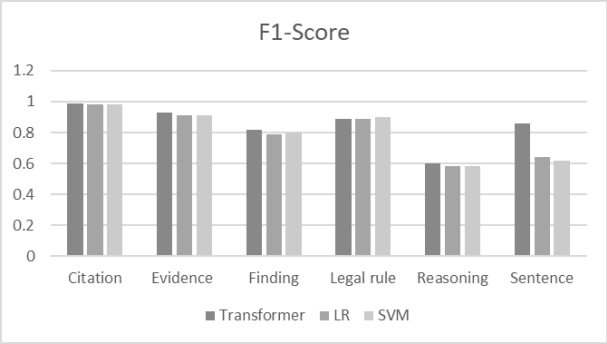


Chart 3: Comparison of F1-Score between various AI techniques




## 5. Conclusion and Future Work

Our approach of using the bert-large-cased model for legal sentence classification has led to noticeable improvements in accuracy and reductions in error over previous approaches in most cases, but not all. Precision is important in increasing confidence in usage of ML based automation as an assistance to humans. However, lack of large amount of data has limited further improvement in accuracy. Also, the improvement, while noticeable, is relatively small, so the benefits need to be evaluated in context of additional cost due to transformers. We did not measure the computation cost of transformer-based classification vs other approaches, however there is enough general data to indicate that transformers are very compute-intensive [10]. This paper provides a valuable data-point in the case of legal sentence classification and comparison of results using bert-large-cased with other approaches. These data points will enable future research work to be more focused.

This work can be extended with a larger dataset if available. It can also be tried with newer and larger transformer models. Given the limited data, few-shot-classification approach is an approach we intend to evaluate, whereby only a very small number of examples of sentences along with their type are used to classify all the sentences in the data. We could not attempt few-shot approach due to the lack of availability of high-power compute machines.


## ACKNOWLEDGMENTS

We thank Dr. Krishnan Pillaipakkamnatt for his support and guidance.


## REFERENCES


[1] Vern R. Walker, Krishnan Pillaipakkamnatt et al. 2019. *Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning*. ASAIL, Montreal, Canada.

[2] Board of Veterans' Appeals of the U.S. Department of Veterans Affairs. 2019. *VetClaims-JSON* https://github.com/LLTLab/VetClaims-JSON

[3] Jack Mumford, Katie Atkinson, and Trevor Bench-Capon. 2021. Machine Learning and Legal Argument. University of Liverpool. https://livrepository.liverpool.ac.uk/3135748/1/CMNA_2021.pdf

[4] HuggingFace. 2019. *bert-large-cased* https://huggingface.co/bert-large-cased

[5] Schakel, A., Beckhaus, G. and Treichl, L. 2023. *Machine learning in the legal industry - potential, pitfalls and how to make it work in real life, Lexology*. Available at: https://www.lexology.com/library/detail.aspx?g=ae37792e-eea3-40a1-9d6d-e764444c3fdf.

[6] Turing. 2022. *The Ultimate Guide to Transformer Deep Learning*. Available at: https://www.turing.com/kb/brief-introduction-to-transformers-and-their-power.

[7] E. de Maat, K. Krabben and R. Winkels. 2010. *Machine Learning versus Knowledge Based Classification of Legal Texts*. In Proceedings of the 2010 Conference on Legal Knowledge and Information Systems (JURIX 2010), 87-96.

[8] Haiyi Zhang, Di Li. *Naïve Bayes Text Classifier*. 2007. IEEE International Conference on Granular Computing (GRC 2007)

[9] M. Ikonomakis, S. Kotsiantis, V. Tampakas. 2005. *Text classification using machine learning techniques*. WSEAS transactions on computers, 4 (8)

[10] Qilu Jiao. *A Brief Survey of Text Classification Methods*. 2023. IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)

[11] aiTechPark. 2023. *Bloomberg law introduces AI-powered solution*. Available at: https://ai-techpark.com/bloomberg-law-introduces-ai-powered-solution

[12] Strubell, E., Ganesh A., and McCallum, A. "*Energy and Policy Considerations for Deep Learning in NLP".* Association for Computational Linguistics (ACL). 2019