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

doi:10.3233/FAIA200870

Identifying the Factors of Suspicion

Morgan A. Gray,[a1] Wesley M. Oliver,[b] and Arthur Crivella [b]
[a] *Crivella Technologies, Ltd*
[b] *Duquesne University School of Law*

Abstract. Probable cause determinations are problematic. Like all court decisions using totality-of-the-circumstances tests, it is difficult to use one decision – or even a few – to foresee a subsequent outcome. No human is capable of reading all the relevant Fourth Amendment opinions relevant to resolving any search and seizure issue. Machines may be capable of this task and to do so they will need to be able to identify particular types of suspicious factors from the various ways courts describe the factors. This project examines the ability of three machine learning models to examine the relevant text of opinions to identify the suspicious factors courts used to determine whether adequate suspicion existed from an intrusion protected by the Fourth Amendment.

Keywords. "information retrieval," "reasonable suspicion," "totality of the circumstances," "*k*-nearest neighbor," "decision tree," "logistic regression"

1. Introduction

Totality-of-the-circumstances legal tests, such as probable cause and reasonable suspicion, do not provide much meaningful guidance for the judges and police officers who have to apply them on a daily basis. Thousands of decisions in the United States, rendered by judges at every place in the judicial hierarchy, have interpreted these legal standards [1]. Machines are certainly capable of digesting far more information than humans, and thus, theoretically have greater capacity to apply judicial interpretations of these standards to subsequent cases. No human could read all of the Fourth Amendment cases courts have decided. And certainly, no human could determine the extent to which courts collectively conclude that a suspicious factor, or a combination of suspicious factors, demonstrate the existence of a current or past crime. The question, then, is whether machines can perform these tasks.

This paper considers the ability of computers to perform the essential first step in evaluating the capacity of computers to evaluate suspicion: examine the text of judicial opinions and identify the type of suspicious circumstances described by the facts. Applying the work of Jaromir Savelka, Huihui Xu, and Kevin Ashley [2] to a new type of dataset, three machine learning models were used to identify the language used to describe various bases of suspicion from judicial decisions which analyzed whether an officer had reasonable suspicion to detain a motorist in order to deploy a drug-sniffing dog. Drug interdiction decisions were evaluated because they typically involve very similar facts – a car is stopped for an ordinary traffic offense and the officer looks for a

[1] Morgan. A. Gray, Corresponding author, Crivella Technologies, Ltd., 3945 Forbes Ave, Pittsburgh, Pennsylvania, United States, 15260, United States of America; E-mail: morgangray99@icloud.com

basis to hold the car until a drug-sniffing dog can arrive – and provide a limited universe of potential bases of suspicion.

The models performed with varying degrees of accuracy, but with 56% accuracy, the logistic regression classifier was able to correctly identify the language of a court as fitting within one of 14 categories. This degree of accuracy suggests that machines are capable of at least the first step required to evaluate reasonable suspicion from a corpus of Fourth Amendment decisions.

2. The Data Set

The data set is comprised of 156 opinions which reflect relevant cases from almost every jurisdiction across the United States. We had success identifying and using a wide array of cases from a profoundly diverse group of jurisdictions. Search and seizure law is defined by the Fourth Amendment to the U.S. Constitution and the standard for permitting the detention of a car for a drug dog is therefore the same in all federal and state courts. If an officer, in any jurisdiction in the United States, has reasonable suspicion to believe drugs are present in a car stopped for a traffic violation, the officer may detain the car for a reasonable amount of time until a drug dog arrives [3].

The opinions were chosen because they all addressed a single legal question – whether an officer had reasonable suspicion to believe a stopped motorist possessed drugs. The cases fitting this criteria were read by one of the attorneys on our team. Each sentence describing a factor of suspicion was annotated to reflect which of twelve categories of suspicion were being described. Additionally, sentences reflecting the court's conclusions about the sufficiency of suspicion were annotated as fitting into one of two categories – a judicial finding of sufficient or insufficient suspicion. Within the 156 cases, 658 separate sentences that described the officer's suspicion were labeled as fitting into one or more of the following categories:

The fourteen categories of suspicion were described as such:

(1) Drug City (DC): travel to or from a city known as a source or endpoint for drugs.
(2) Items in Vehicle (IV): content such as multiple cell phones or signs of long travel.
(3) Masking Agents (MA): including deodorizers, air fresheners, cigars, etc.
(4) Nervousness (N): including nervous behavior such as shaking or trembling.
(5) No Value (NV): any descriptions of officers that legally do not support suspicion.
(6) Prior Convictions (PC): priors involving drug offenses.
(7) Rental (R): motorist driving a rental car.
(8) Suspicious Answers (SA): suspicion resulting from field interrogation.
(9) Suspicious Behavior (SB): actions of defendant based on observation.
(10) Suspicious Circumstances (SC): unusual items, covering a range of possibilities from clothing to music choice.
(11) Suspicious Movements (SM): efforts believed to conceal drugs or a weapon.
(12) Travel Plans (TP): usual routes or inconsistent travel stories.
(13) Suspicion Not Present (SNP): court concluded reasonable suspicion absent.
(14) Suspicion Present (SP): court concluded reasonable suspicion present.

3. Results & Discussion

We used the method developed by Savelka, Ashley, and Xu to evaluate the ability of three machine learning models – decision tree, *k*-nearest neighbor, and logistic regression classifiers – to assess the ability of potential sorting of the language of Fourth Amendment opinions into categories identified by lawyers[2]. Salvelka et al. looked at the ability of these three models to identify sentences in judicial opinions that had been identified by human lawyers to be relevant to interpreting terms in statutes. Our experiment applied this methodology to a new application – the identification of the types of suspicion described by a court in a Fourth Amendment opinion. We applied each classifier to the annotated text of judicial opinions, testing their respective abilities to identify which of fourteen categories, identified in the previous section, was being described by sentences in the opinion.

Computers generally have difficulty classifying sentence-length texts[2][4], but our work with these three classifiers suggests that at least some categories of suspicion can be extracted from judicial language. The logistic regression model performed the best of the three across all categories, with an accuracy rate of 56%, demonstrating a particular aptitude for identifying those sentences describing drug cities, nervousness, prior convictions, masking agents, and rental agreements. The regression classifier was also quite adept at identifying the court's conclusion about the adequacy of suspicion. Figure 1 demonstrates the results for each of the categories for *k*-nearest neighbor and logistic regression classifiers.

NEAREST NEIGHBORS

	precision	recall	f1-score
DC	0.00	0.00	0.00
IV	0.23	0.30	0.26
MA	0.25	0.20	0.22
N	0.18	0.93	0.30
NV	0.00	0.00	0.00
PC	0.20	0.07	0.11
R	0.17	0.60	0.26
SA	0.20	0.12	0.15
SB	0.20	0.09	0.13
SC	0.27	0.12	0.16
SM	1.00	0.14	0.25
SNP	0.14	0.17	0.15
SP	0.71	0.19	0.29
TP	0.00	0.00	0.00
accuracy			0.21
macro avg	0.25	0.21	0.16
weighted avg	0.30	0.21	0.18

LOGISTIC REGRESSION

	precision	recall	f1-score
DC	0.67	0.67	0.67
IV	0.33	0.20	0.25
MA	0.75	0.60	0.67
N	0.38	0.87	0.53
NV	0.00	0.00	0.00
PC	0.58	0.50	0.54
R	0.67	0.80	0.73
SA	0.17	0.25	0.20
SB	0.00	0.00	0.00
SC	0.41	0.59	0.48
SM	1.00	0.14	0.25
SNP	1.00	0.17	0.29
SP	0.77	0.85	0.81
TP	1.00	0.06	0.11
accuracy			0.48
macro avg	0.55	0.41	0.39
weighted avg	0.56	0.48	0.44

Figure 1. We did not include the results from derived from the decision tree classifier because most factors were poorly developed. Only five factors scored above 0.00, but scored fairly well. N scored (P – 0.69, R – 0.73, F-1 – 0.71), R scored (P – 0.50, R – 0.40, F-1 – 0.44), SC scored (P – 0.27, R – 0.88, F-1 – 0.41), SNP scored (P – 1.00, R – 0.50, F-1 – 0.67), SP scored (P – 0.76, R – 0.70, F-1 – 0.73).

The broad range of performance across each category is likely explained by the variety of language that could be used to describe suspicious factors within each category. Predictably, categories that potentially encompass a variety of circumstances, that can be described in a number of ways, were harder for the model to identify than categories that are most often described using very similar terms.

Courts use a fairly common vocabulary to describe masking agents (MA), nervousness (N), and prior convictions (PC), likely explaining the comparative advantage in identifying sentences describing these categories. Categories identifying a broader range of behavior, as one would expect, were comparatively difficult for the

model to identify. Suspicious answers (SA) to questions could take a number of forms. Suspicious behaviors (SB) could include anything from talkativeness to sullenness to combativeness. One would therefore also expect the model to have difficulty identifying sentences describing suspicious circumstances (SC), a category that could include anything from wearing a tie-dye shirt to playing loud religious music. The classifier's comparative success rate (*precision (P):* 0.41; *recall (R):* 0.59) is somewhat remarkable, though likely explained by presence of this category of suspicion was quite common, accounting for 22% of the sentences analyzed.

The regression classifier had remarkable difficulty in identifying factors described by officers as suspicious but which courts disregard as having no legal value. (*P*: 0.00; *R:* 0.00). These cases were coded as having no legal value (NV). As in other catch-all categories, sentences coded in this category could describe an expansive range of circumstances. The performance of the model here was non-existent. This is almost certainly explained by the fact that only two sentences in the dataset were coded NV.

For most categories, the regression model produced similar results for recall and precision, with three notable exceptions – travel plans (*P:* 1.00; *R:* 0.06), suspicious movements (*P*: 1.00; *R*: 0.14), and judicial findings that there was inadequate suspicion (*P*: 1.00, *R*: 0.17). Descriptions of unusual or inconsistent travel plans could be described in any number of ways that are obvious to human readers but may difficult for models to detect, for instance the driver and passenger identifying travel two cities in different parts of the country. Suspicious movements accounted for a small fraction of the sentences annotated. The model's inability to recall judges' conclusions of inadequate suspicion is a mystery. There was no dearth of data for this category, and around 35% of cases analyzed concluded that suspicion was lacking. Yet, the model proved particularly good at identifying conclusions that suspicion was sufficient for a detention (*P:* 0.77; *R:* 0.85).

These results, using models that have not been tailored to this particular data set, suggest that machine learning models can be trained to meaningfully identify particular factors of suspicion in judicial opinions. The logistic regression classifier performed reasonably well across all categories, and exceptionally well in identifying language describing some categories. Moving forward we look to expand our data set and experiment on other models and to adapt them to better serve our categories and data. With larger datasets, a larger number of more precisely defined categories of suspicion can be employed. Improvement is certainly important and necessary, but our results appear to demonstrate the ability of a machine learning to meaningfully identify the types of suspicion a court relied upon on rendering a Fourth Amendment decision.

Acknowledgements

The authors acknowledge the contribution of Wayne Ackman for his efforts in helping to compile the dataset. The authors are also indebted to Dr. Kevin Ashley for his mentorship generally and with this project particularly.

References

[1] Brent E. Newton. The Real World Fourth Amendment. Hastings Const. Law Q. 2016 Apr; 43(4): 759-810.
[2] Jaromir Savelka, Huihui Xu, and Kevin D. Ashley. Improving Sentence Retrieval from Case Law for Statutory Interpretation. In: Floris Bex, editor. Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19), 2019 June 17-21, Montreal, QC, Canada. New York (NY): ACM; c2019. p. 113-112.
[3] Wayne R. LaFave. The "Routine Traffic Stop" from Start to Finish: Too Much "Routine," Not Enough Fourth Amendment, Michigan Law Review. 2004 Aug; 102(8): 1846-1902.
[4] Vanessa G Murdock. Aspects of sentence retrieval. ACM SIGIR Forum. 2001 Dec; 41(2): 127.