October 2020

Towards transparent human-in-the-loop classification of fraudulent web shops

Daphne ODEKERKEN [a,b] and Floris BEX [a,c]
[a] *Department of Information and Computing Sciences, Utrecht University*
[b] *National Police Lab AI, Netherlands Police*
[c] *Tilburg Institute for Law, Technology and Society, Tilburg University*

Abstract. We propose an agent architecture for transparent human-in-the-loop classification. By combining dynamic argumentation with legal case-based reasoning, we create an agent that is able to explain its decisions at various levels of detail and adapts to new situations. It keeps the human analyst in the loop by presenting suggestions for corrections that may change the factors on which the current decision is based and by enabling the analyst to add new factors. We are currently implementing the agent for classification of fraudulent web shops at the Dutch Police.

Keywords. law enforcement, dynamic argumentation, legal case-based reasoning

1. Introduction

Every year, the Dutch police receives thousands of complaints on online trade fraud, many of which concern reports on web shops that do not deliver goods. Nonetheless, not each of these shops has bad intentions: in many cases, the customer fell victim to malfunctioning delivery service, rather than fraud. The Dutch police has a national centre for counteracting online trade fraud, where analysts manually check suspicious web shops. This is a combination of routine work (that could be automated) and more detailed investigation (that should be done by humans). Given the high number of suspicious web shops and the necessity to act quickly, the police experiments with using artificial intelligence (AI) agents to speed up the process. In this paper, we introduce a new agent architecture for web shop classification that relies on static and dynamic algorithms for both rule-based and case-based reasoning. On first thought, this may seem an overly complex solution, given that classification problems are often solved in machine learning by training a model on a labelled data set. However, this classical machine learning approach does not suffice for classification problems in law enforcement, for three reasons.

Handling a dynamic environment. Recently, Wabeke et al. [5] presented their multiyear effort in detecting and removing counterfeit web shops from the .nl DNS zone - a problem similar to ours. They developed two detection systems. Interestingly, one of the claims in their paper is that the makers of counterfeit web shops adapted to their first system. This makes clear that a web shop classifier should be able to adapt to its environment. This issue can be handled by frequently updating the model, but that would require continuous effort from an AI expert. Instead, we aim for a more future-proof solution that directly takes input from analysts into account, as we will show in Section 3.

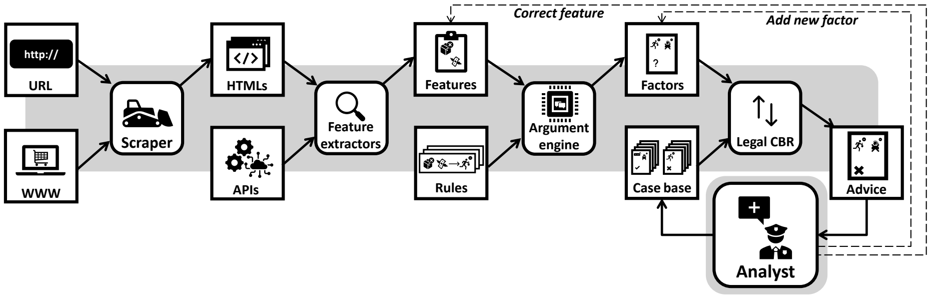

Figure 1. Web shop checker architecture.

Human in the loop. The classification outcome has serious implications: web shops classified as mala fide will be taken offline, while web shops classified as bona fide can be placed on a white list for fraud intake (see [3]). In view of this, it is required that a human analyst checks each advice given by the classifier. However, we should be alert to the control problem [6], i.e. the situation that a human analyst devolves too much responsibility to the classifier and fails to detect cases where the classifier is wrong. To prevent this, the analyst should be kept actively in the loop: he or she should for example be notified of possible mistakes by the classifier and be encouraged to check these situations. An additional motivation for a human-in-the-loop approach is that some factors influencing the decision can only be found by a manual investigation, for example since they require making a payment. In cases where these factors could be relevant, the analyst should be invited to investigate these factors and return the resulting information to the agent.

Transparent. Currently, the result of a web shop check by human analysts is a well-founded advice that includes the factors that made them decide on their conclusion. This is required for various purposes, e.g. alerting citizens and informing the registrar in a web site take-down request, so our agent should be able to produce similar explanations. In general, transparency is one of the key requirements for trustworthy AI applications, as identified by the European Commission's High-Level Expert Group on AI[1].

2. Web shop classification agent architecture: from URL to initial advice

Our proposed agent architecture is illustrated in Figure 1. The initial input is a URL that is considered suspicious. As a first step, a **web scraper** scrapes the web shop's HTML pages. Subsequently, features are extracted from the HTML by **feature extractors**. Some feature extractors require **API** calls to obtain additional information from external organisations. The resulting feature vector is the input for the argumentation engine.

The **argumentation engine** uses a set of defeasible rules to find arguments for or against factors that influence the decision if a web shop should be trusted. Factors are either bona fide (e.g. *"uses https"*) or mala fide (e.g. *"uses fake hallmark logo"*) and are identified by consulting analysts. We use an ASPIC$^+$ [4] implementation that applies rules to features, thus obtaining arguments that support and attack factors; see Figure 2 for an example. Given a set of arguments and the attack relation between them, the argumentation engine determines the set of acceptable arguments by computing the grounded

[1] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines

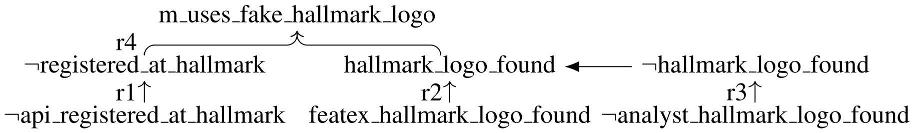
m_uses_fake_hallmark_logo
r4
¬registered_at_hallmark
hallmark_logo_found
¬hallmark_logo_found
r1↑
r2↑
r3↑
¬api_registered_at_hallmark
featex_hallmark_logo_found
¬analyst_hallmark_logo_found

Figure 2. Excerpt from the rule set. If a feature extractor found a hallmark logo at the web site, but an API call returns that this site is not registered at the hallmark company, there is an argument for the mala fide factor "uses fake hallmark logo". Rule r3 is stronger than r2, so if an analyst could not find the logo, then the argument for m_uses_fake_hallmark is attacked on hallmark_logo_found - and removed from the grounded extension.

extension [1]. The grounded extension identifies arguments that can reasonably be accepted; hence all factors for which there is an argument in the grounded extension can reasonably be taken into account in the final decision. The output of the argumentation engine is the set of factors for which there is an argument in the grounded extension.

Finally, the **legal case-based reasoning (CBR)** module compares the factors of the tested web shop to a case base of earlier ⟨factor set, conclusion⟩ pairs. It identifies precedential constraints [2] based on a fortiori reasoning: a web shop is constrained to be mala fide if its factors are at least as "bad" as those of a precedent case labelled mala fide (since all mala fide factors of the precedent case apply to our web shop and all bona fide factors of our web shop apply to the precedent case). Similarly, our web shop is constrained to be bona fide if its factors are at least as "good" as those of a precedent case labelled bona fide. If no precedential constraint applies, the tested web shop is labelled undecided. This way, we obtain an initial advice (bona fide, mala fide or undecided) for our web shop.

3. Interaction between the agent and the analyst in the loop

As shown in Figure 1, the human analyst interacts with the agent in four different ways.
Explanation. The agent explains its initial advice to the analyst. This explanation consists of the factors corresponding to the web shop, together with a precedent case from the case base for which a precedential constraint applies, see Figure 3. If required, a more detailed explanation can be constructed by generating the arguments for these factors.
Correcting features. We define the rule set in such a way that the analyst can overrule feature extractors in stating that a feature is present or absent. Such a correction could lead to a change in the present factors, which may influence the advice. By using a variation on the stability algorithm from [3], we can identify which features can still be obtained by an analyst check and would result in a factor change that alters the advice. These features are presented as suggestions to the analyst, as shown in Figure 3.
Adding factors. Alternatively, the analyst may not agree with the advice since some factor is missing. In that case, he or she can add this factor to the case. This information is stored, so that the analyst implicitly constructs a data set that can eventually be used by an AI expert to develop new feature extractors and argumentation rules for this factor.
Case base update. For web shops that cannot be assigned bona fide or mala fide by some precedential constraint, the advice will be undecided. In this case, the analyst chooses between bona fide and mala fide (based on the factors) and adds this new case to the case base. Note that this cannot cause inconsistencies in the case base, since no precedential constraint applied before. Thanks to these continuous updates of the case base, the agent will be able to classify more web shops as bona fide or mala fide in the future.

Based on automatically extracted information, the web shop www.suspicious-shop.com seems to be bona fide. This advice is based on following factors:

- The Chamber of Commerce number mentioned on the web site exists;
- The VAT number on the web site is valid.

This advice is based on a comparable advice for the web shop www.bona-fide-shop.com. However, the following information would change the advice:

- Payments are transferred to a foreign bank account.
  This mala fide factor can be obtained by making a payment.

Figure 3. Example of explanation for a new advice, based on the factors of an old advice.

4. Discussion and conclusion

We proposed an agent architecture for transparent human-in-the-loop classification that combines dynamic structured argumentation with legal case-based reasoning. This way, it can explain its decisions by the contributing factors and previous cases. Thanks to continuous updates on the case base, it adapts to new situations. Finally, it keeps the human analyst actively involved in the loop by presenting suggestions for analyst checks and enabling the analyst to add new factors that can change the classification outcome.

This agent is currently being implemented for the classification of fraudulent web shops at the Dutch Police. In order to efficiently estimate which features and factors could still change the advice, we work on an extension of our stability algorithm [3].

The implementation of the proposed system requires a significant amount of knowledge engineering, since the rules for the argumentation engine are identified manually. We consider this effort to be more than worthwhile, since the rule set provides a way to generate human-readable explanations; furthermore, it is required to run algorithms for dynamic argumentation [3]. Finally note that many rules can be obtained easily since they fit in a certain scheme - for example, some feature is present if it is detected by a feature extractor or if it is observed by an analyst, see the rules for hallmark_logo_found in Figure 2. In case of conflict, the rule based on the analyst's observation is stronger.

Although the agent architecture is designed for the law enforcement domain, it could also be used for transparent human-in-the-loop classification in other domains - provided that one can identify factors that correspond to one of the two classes. Finally, we only used binary factors, but we plan to extend our approach towards dimensions.

References

[1] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[2] John F Horty. Rules and reasons in the theory of precedent. *Legal Theory*, 17:1–33, 2011.

[3] Daphne Odekerken, AnneMarie Borg, and Floris Bex. Estimating stability for efficient argument-based inquiry. In *Proceedings of the 8th International Conference on Computational Models of Argument*, 2020.

[4] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124, 2010.

[5] Thymen Wabeke, Giovane Moura, Nanneke Franken, and Cristian Hesselman. Counterfighting counterfeit: detecting and taking down fraudulent webshops at a ccTLD. In *Proceedings of the 21st International Conference on Passive and Active Network Measurement*, pages 158–174. Springer, 2020.

[6] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. Algorithmic decision-making and the control problem. *Minds and Machines*, 29(4):555–578, 2019.