Detecting and Explaining Unfairness in Consumer Contracts with Memory Networks

Federico Ruggeri, Francesca Lagioia, Marco Lippi, Paolo Torroni

***Abstract*—We present an approach based on Memory-Augmented Neural Networks for the task of detecting clauses in online Terms of Service that are potentially unfair for the consumer, with the additional goal to explain the legal rationale behind unfairness. The proposed approach goes in the direction of developing neural models for the legal domain, whose output is interpretable and explainable by humans.**

I. Introduction

Terms of Service (ToS) are contracts governing the relations between service providers and users, establishing mutual rights and obligations. Be it for shortage of time or information overload, frequently consumers do not read or understand such contracts [8], unwittingly subscribing to clauses that too often are potentially or clearly unfair [6, 7]. Moreover, even if consumers did read the ToS thoroughly, they would have no means to influence their content. Despite substantive law in place to safeguard their rights by preventing unfair commercial practices, and despite the competence of enforcers for abstract control, consumers and their organizations often lack the legal knowledge or resources needed to take legal action. The risk of individuals becoming overpowered is particularly serious. Artificial intelligence tools have recently been proposed [5] to aid consumer protection organizations and leverage consumers empowerment.

In this paper we present recent advancements[1] of a project, named CLAUDETTE,[2] aimed at using machine learning and natural language processing methodologies to build an intelligent system for the automatic identification of (potentially) unfair clauses in ToS [4]. A system prototype, deployed as a web service, can already be used by consumers who can submit service agreements in a text box and visualise CLAUDETTE's predictions about eight categories of potentially unfair clauses.[3] One limit of such a system, shared by many data-driven machine learning applications, is the lack of transparency and explainability of the outcomes of the system. However, in this context, producing relevant explanations for the unfairness of a clause is crucial, no less than making accurate unfairness predictions [1]. In this work, we illustrate a recent advancement of our methodology, that consists in the introduction of *explanations* to the output of CLAUDETTE. To this end, we exploit Memory-Augmented Neural Networks (MANNs) [9], an architecture that combines the successful learning strategies developed in the machine learning literature for inference with a memory component that can be read and written to.

The key idea behind our approach is that useful explanations may be given in terms of *rationales*, i.e. ad-hoc justifications provided by legal experts, motivating their conclusion to consider a given clause as unfair. A MANN classifier can be trained to identify unfair clauses by using as facts the rationales behind unfairness labels. In this way, a possible explanation of an unfairness prediction can be derived from the list of memories, i.e., the rationales, used by the MANN.

We tested our approach on a collection of 100 ToS been annotated by legal experts following the criteria described in [4]. For this work we selected five unfairness categories and listed all the possible associated rationales, described in the form of self-contained English sentences. We then fed such rationales to the external memory of our MANN classifier. This approach matches or outperforms state-of-the-art classifiers, with the additional feature of producing explanations that are interpretable by non-experts.

II. Legal Rationales of Unfairness

To distinguish and classify instances as fair or unfair, domain experts and decision makers usually rely on their capacity to interpret and apply the relevant legal instruments, trained on their experience with relevant examples. They are also able to provide explanations for their intuitions of unfairness, appealing to standards, rules and principles, possibly expressed by cases, and most significantly by judicial precedents. Encoding such expert knowledge to provide benefit for a consumer is a challenging task.

For the purpose of this study we have conducted an in-depth analysis of the unfair clauses within the Claudette data set and we have created a novel structured corpus of different legal rationales, with regard to the following unfairness categories: liability exclusion (`ltd`), unilateral change (`ch`), termination (`ter`), content removal (`cr`), and arbitration (`a`). The analysis produced a total of 18 rationales for `ltd` (18), 17 for `cr` (17), 28 for `ter` (28), 8 for `ch` and 8 for `a`. Because a single potentially unfair clause can be linked with multiple explanations, the mapping from clauses to rationales included in the KB is one–to–many.

As an example, consider the following `ltd` potentially unfair clause from the Duolingo ToS:

FR and PT are with DISI – University of Bologna. Email: federico.ruggeri6@unibo.it, paolo.torroni@unibo.it. FL is with Law Department at European University Institute. Email: francesca.lagioia@eui.eu. ML is with DISMI – University of Modena and Reggio Emilia. Email: marco.lippi@unimore.it

[1]An extended version of this paper has been submitted for publication in a peer-reviewed journal.

[2]http://claudette.eui.eu/

[3]These include: (i) jurisdiction, (ii) choice of law, (iii) limitation of liability, (iv) unilateral change, (v) unilateral termination, (vi), arbitration clause, (vii) contract by using, (viii) content removal.

> "In the event that Duolingo suspends or terminates your use of the Service or these Terms and Conditions or you close your account voluntarily, you understand and agree that you will receive no refund or exchange of any kind, including for any unused virtual currency or other Virtual Item, any Content or data associated with your use of the Service, or for anything else."

The clause limits the provider's liability by compensation amount consumers may receive and by causes of potential damages, i.e., for disturbances in the availability and reliability of the service (interruption and termination), excluding any guarantees with regard to its provision. Its unfairness can thus be explained by two rationales:

> **[amount]**: the compensation for liability or aggregate liability is limited to, or should not exceed, a certain total amount, or that the sole remedy is to stop using the service and cancel the account, or that you can't recover any damages or losses.
>
> **[discontinuance]**: the provider is not liable for any technical problems, failure, inability to use the service, suspension, disruption, modification, discontinuance, relyability, unavailability of service, any unilateral change, unilateral termination, unilateral limitation including limits on certain features and services or restricttion to access to parts or all of the Service without notice.

Future work includes investigation of how these types of legal rationales are linked to different types of market sectors.

III. Method

The task of unfair clause detection in consumer contracts is formulated as a binary classification problem, in which the model has also access to an external KB containing legal rationales depicting the possible motivations behind a certain type of unfairness.

Formally, a MANN is an architecture coupling a neural model with an external supporting memory [10, 9, 2]. Such a memory brings two important benefits to model representational capabilities:

(1) the memory can act as an auxiliary tool to handle complex reasoning such as capturing long-term dependencies;
(2) the memory can be employed to inject external domain knowledge directly into the model for different purposes, mainly interpretability, transfer learning and context conditioning.

Our approach is centred on the latter advantage and builds on and extends a first experimental setup of MANNs for unfairness detection [3]. From a technical point of view, the model takes the clause to classify as input, referred as the *query* q, and compares it with each element stored into the memory M, m_i, via a (parametric) similarity operation $s(q, m_i)$. As a result, a set of (normalized) similarity scores w_i are retrieved and used to aggregate memory content into a single summary vector $c = \sum_{i=1}^{|M|} w_i \cdot m_i$. Intuitively, this aggregated result can be thought as a fuzzy representation of the memory M conditioned on the given input query q. Indeed, we are only interested in retrieving relevant memory content, that is, functional to a correct classification of the input clause. Lastly, the retrieved memory content is used to enrich (update) the query in order to ease the classification process.

TABLE I
CLASSIFICATION PERFORMANCE ACCORDING TO MACRO-F1 COMPUTED ON 10-FOLD CROSS-VALIDATION FOR UNFAIR EXAMPLES.

	Categories				
Model	A	CH	CR	LTD	TER
SVM	0.350	**0.673**	0.538	0.636	0.636
CNN	0.361	0.654	0.584	0.627	0.612
LSTM	0.326	0.639	0.498	0.589	0.589
MANN (WS)	0.503	0.670	0.596	0.649	0.664
MANN (SS)	**0.526**	0.665	**0.606**	**0.659**	**0.666**

TABLE II
MEMORY INTERACTION STATISTICS FOR CH REPORTING MEMORY USAGE (U), CORRECT MEMORY USAGE OVER UNFAIR EXAMPLES (C) AND OVER EXAMPLES THAT US MEMORY (CP), TOP-1 RANKING VERSION (CP@1) AND AVERAGE MEMORY USAGE PER SAMPLE (APM). BASELINES ALWAYS SELECT THE MOST (@1) OR SECOND MOST (@2) FREQUENT RATIONALE.

Model	**U**	**C**	**CP**	**CP@1**	**APM**
Baseline@1	1.0	0.837	0.837	0.837	0.143
Baseline@2	1.0	0.212	0.212	0.212	0.143
MANN (WS)	0.526	0.445	0.845	0.210	0.752
MANN (SS)	**0.872**	**0.805**	**0.913**	**0.850**	**0.454**

As an extension of the methodology presented in [3], we train the system by providing specific information on which legal rationales better describe each unfair clause at training time. This approach is formally known as *strong supervision* (SS) [9], and in the case of legal rationales encoded in the memory is implemented as a max margin loss at extraction level. It can be described informally as suggesting higher preference for memory elements, i.e., legal rationales, that are labeled by experts as the motivation why some given clauses should be considered unfair. Experimental results show the effectiveness of SS, leading to improved classification performance, more consistent model interpretability and proper memory usage.

IV. Results

The experimental setting is a direct follow-up of [3], where we consider unfairness categories described in Section II. Models are trained via a repeated 10-fold cross-validation routine for robust estimation and macro-average F_1 score is considered as the evaluation metric. Results in Table I show that even a naive combination of a simple MANN architecture and raw knowledge representation yields an increased performance in all investigated unfair categories over traditional knowledge-agnostic models, such as basic neural baselines and the current state-of-the-art SVM solution [4]. When coupled with SS, MANN models gain added performance compared to the end-to-end training methodology known as *weak supervision* (WS). Most importantly, SS brings several benefits from the memory usage point of view: models learn to properly use the memory for inference, avoiding potential ill-behaved scenarios, such as using all the memory for classification. Table II confirms our intuition that SS is crucial for correctly linking potentially unfair clauses to their correct rationales.

REFERENCES

[1] H. Cramer et al. “The effects of transparency on trust in and acceptance of a content-based art recommender”. In: *User Modeling and User-Adapted Interaction* 18.5 (2008), p. 455.

[2] A. Graves, G. Wayne, and I. Danihelka. “Neural Turing machines”. In: *arXiv preprint arXiv:1410.5401* (2014).

[3] F. Lagioia et al. “Deep Learning for Detecting and Explaining Unfairness in Consumer Contracts”. In: *JURIX 2019*. Vol. 322. IOS Press. 2019, p. 43.

[4] M. Lippi et al. “CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service”. In: *Art. Int. and Law* 27.2 (2019), pp. 117–139.

[5] M. Lippi et al. “The Force Awakens: Artificial Intelligence for Consumer Law”. In: *J. Artif. Intell. Res.* 67 (2020), pp. 169–190.

[6] M. Loos and J. Luzak. “Wanted: a bigger stick. On unfair terms in consumer contracts with online service providers”. In: *Journal of consumer policy* 39.1 (2016), pp. 63–90.

[7] H.-W. Micklitz, P. Pałka, and Y. Panagis. “The empire strikes back: digital control of unfair terms of online services”. In: *Journal of consumer policy* 40.3 (2017), pp. 367–388.

[8] J. A. Obar and A. Oeldorf-Hirsch. “The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services”. In: *Information, Communication & Society* 23.1 (2020), pp. 128–147.

[9] S. Sukhbaatar, J. Weston, R. Fergus, et al. “End-to-end memory networks”. In: *Advances in neural information processing systems*. 2015, pp. 2440–2448.

[10] J. Weston, S. Chopra, and A. Bordes. “Memory networks”. In: *arXiv preprint arXiv:1410.3916* (2014).