# Explaining potentially unfair clauses to the consumer with the CLAUDETTE tool

Rūta Liepiņa
Faculty of Law, Maastricht University
and Law Department, EUI

Federico Ruggeri
DISI, University of Bologna

Francesca Lagioia
CIRSFID, University of Bologna and
Law Department, EUI

Marco Lippi
DISMI, University of Modena and
Reggio Emilia

Kasper Drazewski
BEUC

Paolo Torroni
DISI, University of Bologna

## ABSTRACT

This paper presents the latest developments of the use of memory network models in detecting and explaining unfair terms in online consumer contracts. We extend the CLAUDETTE tool for the detection of potentially unfair clauses in online Terms of Service, by providing to the users the explanations of unfairness (legal rationales) for five different categories: arbitration, unilateral change, content removal, unilateral termination, and limitation of liability.

## KEYWORDS

Memory Networks, Terms of Service, NLP

**ACM Reference Format:**
Rūta Liepiņa, Federico Ruggeri, Francesca Lagioia, Marco Lippi, Kasper Drazewski, and Paolo Torroni. 2020. Explaining potentially unfair clauses to the consumer with the CLAUDETTE tool. In *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop, 24 August 2020, San Diego, US.* , 4 pages.

## 1 INTRODUCTION

Online market practices continuously display power asymmetry towards consumers [12, 23]. Several technical solutions have emerged [5, 17, 18], but the focus has largely been on identifying clauses that might be of interest to consumers, in that way navigating the reader through the extensively long agreements [16]. However, the lack of context and explanation of such clauses, as well as limited enforcement possibilities, have hindered the desired goals in consumer protection.

While there seems to be an agreement that most Terms of Service (ToS) agreements contain clearly or potentially unfair clauses [13, 23], it may be insufficient to know which clauses are unfair without providing context for the consumer [9]. Moreover, for such explanations to eventually lead to effective protection, they must be grounded in the current legal framework in the European Union, i.e. The Unfair Contract Terms Directive 93/13/EEC (the Directive).

In this paper we present one possible solution to increase consumer empowerment through technology based on memory networks. Following earlier studies [8, 11], we have introduced the use of legal rationales as explanations of clause unfairness within the updated CLAUDETTE tool. In Section 2 the paper will explore the need for explanations and illustrate ways in which such explanations can be automatically generated. Sections 3 and 4 present the extended knowledge base of legal rationales and methods behind such integration. Section 5 demonstrates the new features of the tool and examples of what information is provided to the consumers when inquiring about the fairness of their contractual terms.

## 2 THE NEED FOR EXPLANATIONS

The need for explainable results by AI systems has been a viral topic in the regulatory territory [2] and has provoked the interest of many scholars [1, 3, 7, 9, 15, 19]. Main themes of this research include interpretability of results produced by AI systems, transparency of the workings of such systems, and the relationship between explainability and trust of the end-users. In the context of consumer contracts, lack of clear explanations of user rights in the terms and conditions has resulted in uninformed consent and truth obstruction by the companies [20]. To remedy the information imbalance, we designed CLAUDETTE, a tool for the automatic detection of potentially unfair clauses in contracts [11]. However, further explanations of detected clauses were not available to the users.

One method to integrate domain knowledge in machine learning classifiers that has been explored in the AI community is the end-to-end memory network model [21, 22], which allows to perform classification by exploiting an additional, external memory of knowledge. Within this memory we stored a collection of legal rationales provided by legal experts. In consumer contracts, in fact, unfair clauses are linked with legal rationales. The feature of providing the user with rationales of why the particular clause can be considered unfair is seen as an important development of the tool for effective empowerment of consumers [9, 10, 14, 15].

## 3 KNOWLEDGE BASE: LEGAL RATIONALES OF UNFAIRNESS

The original training set for the classification tasks included 100 ToS agreements from the most popular online companies that were double-labelled by legal experts, according to the criteria described in [11]. In addition to comprehensive annotation guidelines based on the Directive, its annex with a list of sample clauses which can be described as unfair, and Court of Justice of the European Union decisions, the project also relied on the individual legal expertise and previous experience of the annotators, e.g., in understanding and applying the relevant legal instruments. Given the legal framework, the project focuses on *unfair terms* as defined in the European

*NLLP @ KDD 2020, August 24th, San Diego, US*
© 2020
.

Union. Encoding of this expert knowledge such that it provides benefit for a consumer is a challenging task. In the previous version of the CLAUDETTE tool, users could copy and paste their service agreements into a text-box and the system automatically detected potentially unfair clauses based on nine unfairness categories.[1]

Creating a knowledge base for the detected clauses is a slightly different task. At this stage, we have chosen five unfairness categories: limitation of liability (<ltd>), unilateral change (<ch>), unilateral termination (<ter>), content removal (<cr>), and arbitration (<a>). The knowledge base consists of the rationales and their unique identifiers that are linked to the unfairness categories. In particular, the following distribution of rationales was created based on the information patterns in the online contracts: <ltd> (18), <cr> (17), <ter> (28), <ch> (8), <a> (8). Note that a single potentially unfair clause can be linked with different explanations.

Consider the following clause taken from the Goodreads ToS and classified as (potentially) unfair under unilateral termination:

> "Goodreads may permanently or temporarily terminate, suspend, or otherwise refuse to permit your access to the Service without notice and liability for any reason, including if in Goodreads' sole determination you violate any provision of this Agreement, or for no reason."

It has been associated to the following three rationales:

> **[any_reason]**: since the clause generally states the contract or access may be terminated for any reason, without cause or leaves room for other reasons which are not specified.
>
> **[breach]**: since the contract or access can be terminated where the user fails to adhere to its terms, or community standards, or the spirit of the ToS or community terms, including inappropriate behaviour, using cheats or other disallowed practices to improve their situation in the service, deriving disallowed profits from the service, or interfering with other users' enjoyment of the service or otherwise puts them at risk, or is investigated under any suspicion of misconduct.
>
> **[no_notice]**: since the clause states that the contract or access may be terminated without notice or simply posting it on the website and/or the trader is not required to observe a reasonable period for termination.

Each of the rationales provides an explanation of a different aspect of the given clause. 'Any reason' rationale is the most common type of 'explanation' that is present in all unfairness categories albeit in slightly different shapes. Blanket phrases such as 'any reason', 'no reason' or 'full discretion' are unlikely to pass the contractual term fairness test under the Directive. Similarly, the 'no notice' rationale, which cover situations where the consumer is expected to regularly check the service online pages to update their knowledge about the changing rights and obligations. It can also be argued that a full termination of services based on an alleged breach of contract is unfair under the Directive, especially in the absence of review mechanisms and/or explanations given to the consumers.

For further illustration, consider the clause from the Oculus ToS, which has been detected as (potentially) unfair for the unilateral change category:

> "We may update or revise these warnings and instructions, so please review them periodically."

Detection of unfairness in this context can be explained by two rationales:

> **[anyreason]**: since the clause states that the provider has the right for unilateral change of the contract/services/goods/features for any reason at its full discretion, at any time
>
> **[justposted]**: since the clause states that the provider has the right for unilateral change of the contract/services/goods/features where the notification of changes is left at a full discretion of the provider, i.e. by simply posting the new terms on their website, with or without a direct notification to the consumer

Similar to the previous example, this company has used a general statement to claim full discretion in updating their terms and conditions. Additionally, they have also limited the notification procedure to only posting the updates online with no further clarifications on whether and how the consumer would be informed. Future work of this project includes investigation of these types of legal rationales that are linked to different types of market sectors.

## 4 METHOD

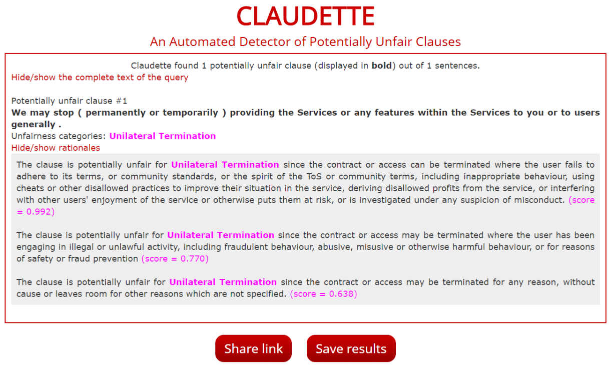
The task of unfair clause detection in consumer contracts is formulated as a binary classification problem, in which the model has also access to an external knowledge base containing legal rationales depicting the possible motivations behind a certain type of unfairness. Formally, an architecture coupling a model with an external supporting memory is known as memory-augmented neural network (MANN) [4, 21, 22]. 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 extends the first experimental setup of MANN's for unfairness detection [8] by considering several categories of legal violations. 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 of as a fuzzy representation of the memory $M$ conditioned on the given input query $q$. Indeed, we are only interested in retrieving memory content that is useful to correctly classify the input clause. Lastly, the retrieved memory content is used to enrich (update) the query in order to ease the classification process. Note that the MANN architecture also allows an iterative interaction with the memory, each time employing the previously updated query, suitable for complex reasoning tasks, such as reading comprehension [6]. However, the task of unfairness

[1] These include the choice of (i) jurisdiction, (ii) choice of law, (iii) limitation of liability, (iv) unilateral change, (v) unilateral termination, (vi), arbitration, (vii) contract by using, (viii) content removal, (ix) privacy included.

detection allows us to limit to a single iteration approach, since it is sufficient to link a single legal rationale to motivate its unfairness.

## 5 DEMO

The CLAUDETTE web service built on the aforementioned MANN-based methodology provides an output such as the one depicted in Figure 1.[2] In particular, the tool offers the user the possibility to enter some text to analyse; the input text is then separated into sentences, and each of them is classified as either unfair or not. In the first case, the system also predicts the unfairness category. For each detected unfair sentence presented in the results web page, CLAUDETTE thus reports the unfairness category and, if any, also the list of legal rationales that were employed by the underlying MANN model during classification, each with a corresponding confidence score. In this way the user is not only informed about the unfairness categories and reasons for unfairness, but also is given an indicator on how relevant these reasons are for the input text.



CLAUDETTE

An Automated Detector of Potentially Unfair Clauses

Claudette found 1 potentially unfair clause (displayed in **bold**) out of 1 sentences.

Hide/show the complete text of the query

Potentially unfair clause #1

**We may stop ( permanently or temporarily ) providing the Services or any features within the Services to you or to users generally .**

Unfairness categories: **Unilateral Termination**

Hide/show rationales

The clause is potentially unfair for **Unilateral Termination** since the contract or access can be terminated where the user fails to adhere to its terms, or community standards, or the spirit of the ToS or community terms, including inappropriate behaviour, using cheats or other disallowed practices to improve their situation in the service, deriving disallowed profits from the service, or interfering with other users' enjoyment of the service or otherwise puts them at risk, or is investigated under any suspicion of misconduct. (score = 0.992)

The clause is potentially unfair for **Unilateral Termination** since the contract or access may be terminated where the user has been engaging in illegal or unlawful activity, including fraudulent behaviour, abusive, misusive or otherwise harmful behaviour, or for reasons of safety or fraud prevention (score = 0.770)

The clause is potentially unfair for **Unilateral Termination** since the contract or access may be terminated for any reason, without cause or leaves room for other reasons which are not specified. (score = 0.638)

**Figure 1: Example of classification performed by the CLAUDETTE tool. Unfair sentences are highlighted in bold and tagged with predicted unfairness label, i.e., category. Additionally, if the memory has been used during classification, the list of exploited legal rationales along with model confidence score (ranging from 0 to 1) is reported.**

Another noteworthy benefit of the use of MANN is the improved detection rates, especially for unfairness categories that have proved harder to identify. An example of limited liability clauses explored in [8], showed how memory network improves upon the state of the art support vector machine approach:

| Model | Precision | Recall | F1 |
|---|---|---|---|
| State of the art SVM | 52.52 | 81.57 | 63.7 |
| Memory Network | 68.36 | 84.31 | 64.33 |

## 6 CONCLUSIONS

This paper presents an extension of the automated detection of unfair terms in consumer contracts by adding explanations through memory network models. It directly addresses the call for more explainable and transparent AI results and furthers the goal of empowering consumers by providing legal rationales on why certain clauses have been detected as potentially unfair, as well as showing the confidence scores of such explanations. In the future, we aim to test different variants of the MANN model to improve the capability of the network to exploit the knowledge, as well as to improve the user experience of the current extension.

We also plan to extend the methodology to privacy policies, which are much more complex documents, for which not only potential unfairness should be checked, but also comprehensiveness and compliance to the existing regulations.[3]

## REFERENCES

[1] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, 2017.

[2] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of ai under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*, 2017.

[3] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[4] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[5] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 531–548, 2018.

[6] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The Goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015.

[7] Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. In *International conference on machine learning*, pages 1378–1387, 2016.

[8] Francesca Lagioia, Federico Ruggeri, Kasper Drazewski, Marco Lippi, Hans-Wolfgang Micklitz, Paolo Torroni, and Giovanni Sartor. Deep learning for detecting and explaining unfairness in consumer contracts. In *Legal Knowledge and Information Systems: JURIX 2019: The Thirty-second Annual Conference*, volume 322, page 43. IOS Press, 2019.

[9] Brian Y Lim, Anind K Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2119–2128, 2009.

[10] Marco Lippi, Giuseppe Contissa, Agnieszka Jablonowska, Francesca Lagioia, Hans-Wolfgang Micklitz, Przemyslaw Palka, Giovanni Sartor, and Paolo Torroni. The force awakens: Artificial intelligence for consumer law. *J. Artif. Intell. Res.*, 67:169–190, 2020.

[11] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139, 2019.

[12] Marco Loos and Joasia Luzak. Wanted: a bigger stick. on unfair terms in consumer contracts with online service providers. *Journal of consumer policy*, 39(1):63–90, 2016.

[13] Hans-W Micklitz. The proposal on consumer rights and the opportunity for a reform of european unfair terms legislation in consumer contracts. 2010.

[14] Bonnie M Muir. Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11):1905–1922, 1994.

[15] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

[16] Jonathan A Obar and Anne Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147, 2020.

[17] Przemysław Pałka and Marco Lippi. Big data analytics, online terms of service and privacy policies. *Research Handbook on Big Data Law edited by Roland Vogl*, 2020.

[2] http://claudette.eui.eu/demo/answers/vYgfZetiN2.html

[3] One of the authors, Francesca Lagioia, has been supported by project "CompuLaw", funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 833647).

[18] Hugo Roy, JC Borchardt, I McGowan, J Stout, and S Azmayesh. Terms of service; didn't read. *Web Page, June. URL https://tosdr. org*, 2012.

[19] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[20] Edith G Smit, Guda Van Noort, and Hilde AM Voorveld. Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in europe. *Computers in Human Behavior*, 32:15–22, 2014.

[21] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

[22] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

[23] Chris Willett. *Fairness in consumer contracts: The case of unfair terms*. Ashgate Publishing, Ltd., 2007.