Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210318

Assessing the Cross-Market Generalization Capability of the CLAUDETTE System

Agnieszka JABLONOWSKA [a], Francesca LAGIOIA [a,b,1], Marco LIPPI [a,c,2], Hans-Wolfgang MICKLITZ [d], Giovanni SARTOR [a,b] and Giacomo TAGIURI [a,e,3]

[a] *Law Department, European University Institute*
[b] *CIRSFID - Alma AI, University of Bologna*
[c] *DISMI, University of Modena and Reggio Emilia*
[d] *Robert Schuman Centre, European University Institute*
[e] *Polish Academy of Science*

Abstract. We present a study aimed at testing the CLAUDETTE system's ability to generalise the concept of unfairness in consumer contracts across diverse market sectors. The data set includes 142 terms of services grouped in five sub-sets: travel and accommodation, games and entertainment, finance and payments, health and well-being, and the more general others. Preliminary results show that the classifier has satisfying performance on all the sectors.

Keywords. Unfair clause detection, machine learning, cross-market analysis

1. Introduction

In the AI and Big data era, where suppliers' power is boosted by technologies, consumer-empowering technologies are needed to support the countervailing power of civil society [1,2,3]. The CLAUDETTE project contributes to this goal, by automating the assessment of standard terms' compliance with EU consumer law. It adopts a supervised machine learning approach, based on a corpus of contracts annotated by domain experts [4], where clauses are labeled either as fair or unfair.

The aim of this work is to study whether and to what extent CLAUDETTE is able to generalize the concept of unfairness across diverse market sectors. To address this question we extended the data set to two additional domains: Health and Finance. The present study allowed us to generate new insights from the pre-existing corpus. In particular, two further groups of companies displaying a degree of sector-specificity, i.e., Travel and Games, were isolated from the original data set.

The paper is organised as follows. In Section 2 we describe the extended corpus, the document annotation procedure, and we briefly introduce a new category of unfair

[1]F. Lagioia and G. Sartor have been supported by the H2020 ERC Project "CompuLaw" (G.A. 833647)
[2]M. Lippi has been supported by the SCUDO project, within the POR-FESR 2014-2020 programme of Regione Toscana.
[3]G. Tagiuri has been supported by the National Science Centre in Poland (G.A. UMO-2019/35/B/HS5/04444)

clauses. We also explain the adopted methodology and discuss the results. Section 3 analyses some possible causes of misclassifications, affecting the CLAUDETTE performance. Finally, Section 4 concludes and presents future research directions.

2. Data set and experimental setting

In our previous works [4,5] we produced a data set consisting of 100 relevant online Terms of Service (ToS) of online platforms, analysed by legal experts and marked in XML. In this research, we increased the data set by adding 42 new contracts, marked by two independent annotators, for a total of 142 ToS.[4] The new documents were selected among those offered by some of the major players in the health and finance market sectors. For the purpose of this study, we split the data set in five groups – i.e., (i) Finance and Payments, (ii) Health and Well-being, (iii) Games and Entertainment, (iv) Travel and Accommodations, and (v) the more general class Others, which contains the remaining contracts originally included in the CLAUDETTE data set. The division of documents into groups is unbalanced: 21 in Finance, 24 in Health, 18 in Games, 8 in Travel, and 71 in others.[5] Such an unbalance is due to two main reasons. On the one hand, the aim pursued in this study led us to collect and analyze new ToS in the Health and Finance domains, which present some peculiarities if compared to other online services. On the other hand, given the effort needed to increase the data set, we decided to reuse the pre-existing corpus.

The annotations reflect the methodology described in [4], where we identified eight different categories of unfair clauses, establishing (1) jurisdiction in a country different than consumer's residence (`<j>`); (2) choice of a foreign law (`<law>`); (3) liability limitations (`<ltd>`); (4) the provider's right to unilaterally terminate the contract/access to the service (`<ter>`); (5) the provider's right to unilaterally modify the contract/the service (`<ch>`); (6) the mandatory arbitration before the court proceedings can commence (`<a>`); (7) the provider's right to unilaterally remove consumer content (`<cr>`); and (8) the consumer consent to the agreement simply by using the service, downloading the app or visiting the website, (`<use>`). In this research we present an additional category of potentially unfair clauses, i.e., those stating (or implicitly assuming) that (9) the scope of consent granted to the ToS incorporates also the privacy policy, which forms part of the "General Agreement" (`<pinc>`). As reported by [6,4,7,8] such categories are widely

[4]The corpus is made freely available for research purposes at the following link: https://claudette.eui.eu/corpus_142_ToS.zip.

[5]In particular, the Finance group includes the following ToS: Bondora, ETFmatic, GoFundMe, Google Payments GB, Google Pay, Kickstarter, Klarna credit agreement, Klarna.com, Ledger.com, Ledger Live, Monzo, Paypal Italy, Revolut, Swanest, Transferwise, Trustly, Visa Solution, Wefox, Western Union for Italy, Xoom, YNAB. The Helath sector includes: 23andme, Ada, Ava, Betterpoints UK, Clue, Endomondo, Fitbit, Flo, Flow, Headspace, Hexoskin, idoc24, iHealth, Kardia, Kry, Lady Cycle, Muse, Mysugr, MyHeritage, Natural Cycles, Polar, Skinvision, Unmind, Woebot. Games includes: ElectronicArts, Epic Games, Habbo, Lindenlab, Masquerade, Nintendo, Oculus,Paradox, Pokemon Go, Rovio, Shazam, Sporcle, Spotify, Steam, Supercell, Ubisoft, World of Warcraft, Zynga. Travel includes: Airbnb, Booking.com, Couchsurfing, eDreams, Expedia, Ryanair, Skyscanner, Verychic. Finally, the Others group includes: 9gag, Academia, Alibaba, Amazon, Atlas, Badoo,Blablacar, Box, Crowdtangle, Dailymotion, Deliveroo, DeviantArt, Diply, Dropbox, Duolingo, eBay, Evernote, Facebook, Foursquare, Garmin, Goodreads, Google, Grammarly, Grindr, Groupon, Happn, HeySuccess, Imgur, Instagram, Lastfm, Match, LinkedIn, Microsoft, Moves, Mozilla, Musically, Myspace, Netflix, Onavo, Opera, Pinterest, Quora, Reddit, Skype, Slack customer, Slack user, Snap, Syncme, Tagged, Terravision, TikTok, Tinder, TripAdvisor, TrueCaller, Tumblr, Twitch, Twitter, Uber, Viber, Vimeo, Vivino, WeChat, Weebly, WeTransfer, WhatsApp, Yahoo, Yelp, YNAB, YouTube, Zalando, Zara, Zoho.

used in ToS for online platforms. To capture the different degrees of (un)fairness we appended a numeric value to each XML tag, with 1 meaning clearly fair, 2 potentially unfair, and 3 clearly unfair, according to the criteria defined in [4,5]. We consider a binary classification task: the positive class is made by all potentially or clearly unfair clauses, for all categories, indiscriminately, and the negative class by all the remaining clauses. We used each of the four sectors, in turn, as a test set, whereas the three remaining sectors, plus all the contracts that belong to none of those four sectors, constituted the training set. For each sector, we performed an additional inner 5-fold cross-validation on the training set to choose the best C hyper-parameter for the linear support vector machine.

Table 1 reports the values of precision (P), recall (R), and F_1 score for each sector considered as test set. P is the percentage of clauses predicted as positive which are really positive (thus accounting for false positives), R is the percentage of correctly detected positive clauses (thus accounting for false negatives) and F_1 is the harmonic mean between P and R. The results show that the classifier has satisfying performance on all the sectors, especially in terms of recall. The sector with the best performance is the Health sector, while the Travel sector has the lowest performance. Both Games and Travel sectors suffer in particular in terms of precision. Section 3 further discusses the reasons why certain types of clauses are wrongly detected as false positives or rather missed by the classifier.

Table 2 shows the percentage of detected potentially unfair clauses (i.e., the recall of the system) for each sector and for each category. To assess the weight of each category across sectors (and therefore on the overall performance), Table 3 shows the average number of clauses per ToS for each sector and category. This information clearly shows which are the most critical clause categories to detect across sectors. The `<pinc>` category (clauses including consent to data processing in the contract, see Section 2) has quite a low recall, especially for Finance and Travel. However, it has to be remarked that at most one clause for such category is usually encountered in a contract. On the other hand, a low recall in all sectors affects limitation of liability clauses, despite their high frequency. Additionally, we can note how unilateral termination and contract by using categories have a low recall in Finance, and arbitration clauses in Travel. More details on the false positives and false negatives are presented in the next section.

Table 1. Experimental results on the cross-sector analysis. For each test sector, we report the micro-averaged precision (P), recall (R), and F_1 as the harmonic mean between the first two metrics.

Sector	P	R	F_1
Finance	0.689	0.739	0.713
Games	0.629	0.849	0.723
Health	0.685	0.809	0.741
Travel	0.597	0.778	0.675

Table 2. Percentage of correctly detected clauses (recall) for each sector and category.

Sector	A	CH	CR	J	LAW	LTD	PINC	TER	USE
Finance	1.000	0.829	0.923	0.765	0.833	0.734	0.375	0.739	0.630
Games	0.833	0.902	0.854	0.920	0.885	0.796	0.769	0.919	0.875
Health	0.950	0.885	0.848	0.909	0.848	0.754	0.786	0.840	0.820
Travel	0.667	0.909	0.875	0.900	0.800	0.633	0.500	0.929	0.847

Table 3. Average number of clauses per contract, for each sector and for each category.

Sector	A	CH	CR	J	LAW	LTD	PINC	TER	USE
Finance	0.333	3.333	0.619	0.810	0.857	7.333	0.381	4.000	1.286
Games	1.333	3.389	2.667	1.389	1.444	6.278	0.722	4.778	2.667
Health	0.833	4.708	1.917	1.375	1.375	8.792	0.583	5.458	2.542
Travel	1.125	1.375	1.000	1.250	1.250	3.750	0.250	1.750	1.625

3. Error analysis

By performing an analysis of the classification errors, we identified three main issues that could cause wrong classifications: (1) the presence of rare linguistic patterns and lexical choices; (2) the specificity of the content of some clauses and (3) the existence of sector-specific regulations.

Linguistic and lexical patterns. Rare linguistic and lexical patterns present in online Terms of Service may be due either to the specificity of a service or to the country in which the service originates (e.g., not in English speaking countries, where most of the ToS in the data set originate). Even though uncommon semantic and lexical formulations can appear in all sectors, we empirically observe that they more frequently emerge in Finance and Health. Since ToS belonging to these sectors were added more recently in the corpus, peculiarities of and changes in recurrent linguistic expressions may also be due to the different times at which the ToS were drafted and analysed.

As an example, consider the following clause taken from the AVA health app ToS (updated May 27th, 2019):

> *Any use of the Products or Site other than as specifically authorized herein, without the prior written permission of Ava is strictly prohibited, and Ava may terminate the license granted herein with immediate effect.*

In this case, "terminate the *license*" slightly differs from the more typical expression "terminate the *contract*". This hypothesis of failure in correctly classifying the clause as `<ter2>` finds support in a similar formulation of a misclassified clause in the Flo ToS (updated February 5th 2020).We plan to examine whether ontologies my help in dealing with such terminological issues.

Content-specificity. The second cause of failure concerns the content–specificity of certain clauses, especially in relation to services dealing with both a digital and a physical component.

While most of the ToS in the original data set concern services that are only digital, others rely on physical devices, are integrated with a physical service, and/or allow to place orders for physical goods. This occurs more frequently in Travel and Health, where, for instance, fitness apps are usually associated to wearable devices. As an example, consider the following clause taken from the Lady Cycle ToS (updated July 8th, 2018):

> *By using Lady Cycle, you acknowledge that you have read and understood the tutorial and manual for its use.*

The clause above has been incorrectly classified by CLAUDETTE as unfair. We can speculate that this is due to the linguistic similarities with the typical consent by using (`<use2`) clauses. Despite these similarities, it rather relates to consumer's acquaintance with product-related instructions. As noted in Section 2, the correct detection of limitation of liability clauses seems to be particularly problematic for all the analysed sectors.

Once again, these may be due to their domain-specific content. Consider for instance, the following clause, not recognized as potentially unfair by CLAUDETTE, and taken from the Hexoskin ToS (updated August 31st, 2019) in the Health sector:

> *We are not responsible for any health problems that may result from training programs, products, or events you learn about through the Hexoskin Services.*

Content-specificity also characterize a number of consent by using clauses in the finance and payments sub-set, where multiple kinds of agreement are mentioned in ToS in relation to multiple services. Consider the following clause taken from the Paypal ToS (updated July 30th, 2021):

> *If we offer you the new checkout solution service and you choose to use it, in addition to this User Agreement, you agree to the following further terms relating to the following capabilities: when you use our APM functionality as part of the new checkout solution, the PayPal Alternative Payment Methods Agreement; and when you use: our Advanced Credit and Debit Card Payments service as part of the new checkout solution; and our Fraud Protection as part of the new checkout solution; Our Fees for using the new checkout solution apply.*

Sector-specific regulations. The third type of cause of misclassification relates to the existence of sector-specific regulations. Such regulations may induce businesses to include in sectorial contracts certain clauses whose content is markedly different from the content of other clauses having a similar wording. For example, the Payment Services Directive II,[6] lays down quantitative limits to losses that the payer may be obliged to bear in case of unauthorised transactions. Specifically, the EUR 50 amount fixed by Art. 74 can recurrently be found in ToS of payment services providers, such as the Transferwise ToS (updated July 28th, 2020):

> *You will be liable for the first 50 EUR of any unauthorised payments if we believe you should have been aware of the loss, theft or unauthorised use.*

This clause concerns the limitation of consumers' liability (thus to their advantage), classified as unfair since its language is similar to clauses stating (unfair) limitation of liabilities of businesses.

Additional remarks. We further analysed the false positives of the Games and Travel sectors, which appear to be the most difficult ones for CLAUDETTE, in terms of specificity.

Regarding Games, we noticed how over 17% of false positives contain the word *arbitration* (or at least its root, like *arbitrate*), and around 16% include the terms *liable* or *liability*. As for arbitration, many of such clauses are sentences that describe the arbitration procedures. For limitation of liability, a significant number of false positives consists in statements confirming providers liabilities, that often use the same terms of potentially unfair clauses excluding such liabilities. One typical example of such clauses is the following one, from Supercell (updated on August 1st, 2017).:

> *Nothing in these terms of service shall affect the statutory right of any consumer or exclude or restrict any liability resulting from gross negligence or willful misconduct of Supercell or for death or personal injury arising from any negligence or fraud of Supercell.*

[6] Directive (EU) 2015/2366 of the European Parliament and of the Council of 25 November 2015 on payment services in the internal market.

Such a clause, stating that the provider is responsible for damages to the consumer, is very similar to unfair clauses stating that the provider is not responsible.

Regarding Travel, around 67% of false positives derive from the Airbnb (46%) and Expedia (21%) ToS. In some cases, such clauses are meant to inform parties on the possible behaviour by a third party, as in the following case taken from Expedia (updated on February 21st, 2018):

Airlines and other travel suppliers may change their prices without notice.

4. Conclusion

Our study investigated the ability of the original CLAUDETTE model to generalize the concept of unfair clauses regarding four market sectors, i.e., games and entertainment, travel and accommodation, finance and payments and health and well-being. The results show that the classifier has satisfying performance on all the sectors, especially in terms of recall. The analysis of false positives and false negatives revealed some possible types of causes of misclassification, including linguistic and lexical patterns, content-specificity, and sector-specific regulations.

In the future, our final goal is to enable CLAUDETTE to automatically detect unfair clauses in consumer contracts across diverse critical market sectors. To this end, we plan to increase the data set with more recent ToS and enlarge the number of market sectors under investigation. We also plan to apply more sophisticated NLP techniques, such as transformers or methods based on sentence embeddings [9]. Finally, we will also compare the current approach to a multi-class formulation, by adding the class of clearly fair clauses for each specific category as an additional output of our system.

References

[1] Lippi M, Contissa G, Lagioia F, Micklitz HW, Pałka P, Sartor G, et al. Consumer protection requires artificial intelligence. Nature machine intelligence. 2019;1(4):168-9.

[2] Lippi M, Contissa G, Jablonowska A, Lagioia F, Micklitz HW, Palka P, et al. The force awakens: Artificial intelligence for consumer law. Journal of artificial intelligence research. 2020;67:169-90.

[3] Thorun C, Diels J. Consumer protection technologies: an investigation into the potentials of new digital technologies for consumer policy. Journal of Consumer Policy. 2020;43(1):177-91.

[4] Lippi M, Pałka P, Contissa G, Lagioia F, Micklitz HW, Sartor G, et al. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law. 2019;27(2):117-39.

[5] Ruggeri F, Lagioia F, Lippi M, Torroni P. Detecting and explaining unfairness in consumer contracts through memory networks. Artificial Intelligence and Law. 2021:1-34.

[6] Loos M, Luzak J. Wanted: a bigger stick. On unfair terms in consumer contracts with online service providers. Journal of consumer policy. 2016;39(1):63-90.

[7] Dari-Mattiacci G, Marotta-Wurgler F. Learning in Standard Form Contracts: Theory and Evidence. NYU Law and Economics Research Paper. 2018;(18-11).

[8] Micklitz HW, Pałka P, Panagis Y. The empire strikes back: digital control of unfair terms of online services. Journal of consumer policy. 2017;40(3):367-88.

[9] Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-Art Natural Language Processing. In: Proc. 2020 EMNLP Conference. Online: Association for Computational Linguistics; 2020. p. 38-45.