ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Pałka, P., Contissa, G., Lagioia, F., Micklitz, H., Sartor, G., Torroni, P., et al. (2019). CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. ARTIFICIAL INTELLIGENCE AND LAW, 27(32), 117-139 [10.1007/s10506-019-09243-2].

Availability:
This version is available at: https://hdl.handle.net/11585/681638 since: 2020-02-03

Published:

DOI: http://doi.org/10.1007/s10506-019-09243-2

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

09 November 2026

This is the final peer-reviewed accepted manuscript of
Lippi, M., Pałka, P., Contissa, G. *et al.* CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artif Intell Law* **27,** 117–139 (2019). https://doi.org/10.1007/s10506-019-09243-2

The final published version is available online at:
https://doi.org/10.1007/s10506-019-09243-2

Rights / License:
The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service

Marco Lippi
marco.lippi@unimore.it
DISMI – University of Modena and Reggio Emilia, Italy

Przemysław Pałka
przemyslaw.palka@yale.edu
Yale Law School Center for Private Law, Information Society Project, New Haven, USA

Giuseppe Contissa
giuseppe.contissa@unibo.it
Francesca Lagioia
francesca.lagioia@eui.eu
CIRSFID, University of Bologna, Italy – Law Department, European University Institute, Villa La Fonte, San Domenico di Fiesole, Florence, Italy

Hans-Wolfgang Micklitz
hans.micklitz@eui.eu
Law Department, European University Institute, Villa La Fonte, Via delle Fontanelle 10, 50014 San Domenico di Fiesole, Florence, Italy

Paolo Torroni
p.torroni@unibo.it
DISI – University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy

Abstract
Terms of service of on-line platforms too often contain clauses that are potentially unfair to the consumer. We present an experimental study where machine learning is employed to automatically detect such potentially unfair clauses. Results show that the proposed system could provide a valuable tool for lawyers and consumers alike.

Keywords
Machine learning · Terms of service · Potentially unfair clauses · Natural language processing

1 Introduction

A recent survey on policy-reading behaviour (Obar and Oeldorf-Hirsch 2016) reveals that consumers rarely read the contracts they are required to accept. This resonates with our direct experience and with what has long been said, that the biggest lie on the Internet is "I have read and agree to the terms and conditions". We use smartphones to gather and share information, connect on social media, entertain ourselves, check our online banking and so on. Virtually every app we install and website we browse have their own Terms of Service (ToS), i.e. contracts governing the relation between providers and users, establishing mutual rights and obligations. Such contracts are also known as "terms and conditions", "service agreements", "statements", or simply "terms". They bind us by the time we switch on the phone or browse a website. However, we are not necessarily aware of what we just agreed upon.

There are reasons why many consumers do not read or understand ToS, as well as privacy policies or end-user license agreements (EULA) (Bakos et al. 2014). Reports indicate that such documents can be overwhelming to the few consumers who actually venture to read them (Department of Commerce 2010). It has been estimated that actually reading the privacy policies alone would carry costs in time of over 200 hours a year per Internet user (McDonald and Cranor 2008). Another problem is that even if consumers did read the ToS thoroughly, they would have no means to influence their content: the choice is to either agree to the terms offered by a web app or simply not use the service at all.

All this created a need for limitations on traders' contractual freedom, not only to protect consumer interests, but also to enhance the consumers' trust in transnational transactions and improve the common market (Nebbia 2007). European consumer law aims to prevent businesses from using so-called "unfair contractual terms" in contracts they unilaterally draft and require consumers to accept (Reich et al. 2014). According to the Unfair Contract Terms Directive (UCTD), a "term" or "clause" (i.e., a sentence, statement on or paragraph expressing a contractual norm that specifies parties rights and obligations) is unfair if, "contrary to the requirement of good faith, it causes a significant imbalance in the parties rights and obligations arising under the contract, to the detriment of the consumer".[1] This definition is supplemented by an Annex containing an "indicative and non-exhaustive list of the terms which may be regarded as unfair" (art. 3.3) and by over 50 ECJ decisions (Micklitz et al. 2017). Law regarding such terms applies also to the ToS of on-line platforms (Loos and Luzak 2016). In spite of it all, such platforms' owners do use in their ToS unfair contractual clauses (Micklitz et al. 2017), notwithstanding European law, and regardless of consumer protection agencies, which have the competence, but not necessarily the resources, to fight against such unlawful practices.

To address this problem, we propose a machine learning-based method and tool for partially automating the detection of potentially unfair clauses (contractual provisions). In particular, we offer a sentence classification system able to detect full sentences, or paragraphs containing potentially unlawful clauses.[2] Such a tool could improve consumers' understanding of what they agree upon by accepting a contract, as well as serve consumer protection organizations and agencies, by making their work more

[1] See the Council Directive 93/13/EEC on Unfair Terms in Consumer Contracts, art. 3.1.

[2] We remark that, from the point of view of natural language processing, we are handling a pure sentence classification task, as we detect full statements and not directly single clauses

effective and efficient, by helping them scan and monitor a large number of documents automatically.

This paper builds upon and significantly extends results presented by Lippi et al. (2017) after a smaller-scale study where a Support Vector Machine (SVM) was trained on a 20-document corpus. With respect to previous work, the contributions of this study are:

• The extension of the corpus, which now consists of 50 contracts (over 12,000 sentences), enabling better training and evaluation of the methods;

• A comparison with several other machine learning systems, including some recent deep learning architectures for text categorization, and a structured SVM for collective classification, which takes into account the sequence of sentences within a document;

• The extension of the classification task from a mere detection of potentially unfair clauses to a more informative classification of such clauses into categories;

• The description of a web server, named CLAUDETTE, which we have made available to the community, so as to allow users to submit query documents and gauge the performance of our methods in autonomy.

The paper is organized as follows. In Sect. 2 we describe the problem from a legal angle. In Sect. 3 we describe the extended corpus and the document annotation procedure. Section 4 explains the machine learning methodology employed in the system, whereas Sect. 5 discusses results. Section 6 describes the web server. Section 7 discusses related work. Section 8 concludes with a look to future research.

2 Problem description

This section provides the necessary background on the European consumer law on unfair contractual terms (clauses). We explain what an unfair contractual term is, present the legal mechanisms created to prevent business from employing unfair terms, and describe our contribution to these mechanisms.

According to art. 3 of the Directive 93/13 on Unfair Terms in Consumer Contracts, a contractual term is unfair if: (1) it has not been individually negotiated; and (2) contrary to the requirement of good faith, it causes a significant imbalance in the parties' rights and obligations, to the detriment of the consumer. This general definition is further specified in the Annex to the Directive, containing "an indicative and non-exhaustive list of the terms which may be regarded as unfair", as well in a few dozen judgments of the Court of Justice of the EU (Micklitz and Reich 2014). Examples of unfair clauses encompass taking jurisdiction away from the consumer, limiting liability for damages on health and/or gross negligence, imposing obligatory arbitration in a country different from consumer's residence etc.

Loos and Luzak (Loos and Luzak 2016) identified five categories of *potentially unfair* clauses often appearing in the terms of on-line services: (1) establishing jurisdiction for disputes in a country different than consumer's residence; (2) choice of a foreign law governing the contract; (3) limitation of liability; (4) the provider's right to unilaterally terminate the contract/access to the service; and (5) the provider's right to unilaterally modify the contract/the service. Our research has identified three additional categories: (6) requiring a consumer to undertake arbitration before the court proceedings can commence; (7) the provider retaining the right to unilaterally remove consumer content from the service, including in-app purchases; (8) having a consumer accept the agreement simply by using the service, not only without reading it, but even without having to click on "I agree/I accept".

The 93/13 Directive creates two mechanisms to prevent the use of unfair contractual terms: *individual* and *abstract* control of fairness. The former takes place when a consumer goes to court: if a court finds that a clauses is unfair (which it can do on its own motion), it will consider that the clause is not binding on the consumer (art. 6). However, most consumers do not take their disputes to courts. That is why abstract fairness control has been created. In each EU Member State, consumer protection organizations have the competence to initiate judicial or administrative proceedings, to obtain the declaration that clauses in consumer contracts are unfair. The national implementations of abstract control differ in various ways. For instance, consumer protection agencies and/or consumer organizations may be involved to a different degree, there may or may not be fines for using unfair contractual terms, etc. (Schulte-Nölke et al. 2008). One thing that all member states have in common is that if a business uses unfair terms in their contracts, in principle there is always a competent party with the authority to challenge such contracts.

Unfortunately, the legal mechanism for enforcing the prohibition of unfair contract terms have failed to effectively counter this practice so far. As reported by some literature (Loos and Luzak 2016), and as our own research indicates (Micklitz et al. 2017), unfair contractual terms are, as of today, widely used in ToS of online platforms.

In our previous research (Micklitz et al. 2017), we developed a theoretical model of tasks that human lawyers currently need to carry out, before starting the legal proceedings concerning the abstract control of fairness of clauses. These include: (1) finding and choosing the documents; (2) mining the documents for potentially unfair

clauses; (3) conducting the actual legal assessment of fairness; (4) drafting the case files and beginning the proceedings. Our work aims to automate the second step, enabling a senior lawyer to focus only on clauses that are found by a machine learning classifier to be potentially unfair, thus saving significant time and labor.

We focus on *potentially* unfair clauses for two reasons. First, we may be unsure whether a certain type of clause falls under the abstract legislative definition of an "unfair contractual term". From a legal standpoint, a given clause can be deemed unfair with absolute certainty only if a competent institution, such as a national court having refereed to the European Court of Justice, has ruled in that sense. That is the case for certain kinds of clauses, such as a jurisdiction clause indicating a country different from the consumer's residence, or limitation of liability for gross negligence (Micklitz et al. 2017). In other cases the unfairness of a clause has to be argued for, showing that it creates an unacceptable imbalance in the parties' rights and obligations. A consumer protection body might want to take the case to a court in order to authoritatively establish the unfairness of that clause, but a legal argument for that needs to be created, and the clause may eventually turn out to be judged fair. Furthermore, unfairness may depend not only on a clause's textual content, but also on the context in which the clause is to be applied. For instance, a mutual right to unilaterally terminate the contract might be fair in some cases, and unfair in others, for example if unilateral termination would entail losing some digital content (purchased apps, email address, etc.) on the side of the consumer.

3 Corpus annotation

The corpus consists of 50 relevant on-line consumer contracts, i.e., ToS of online platforms. Such contracts were selected among those offered by some of the major players in terms of number of users, global relevance, and time the service was established.[3] Such contracts are usually quite detailed in content, are frequently updated to reflect changes both in the service and in the applicable law, and are often available in different versions for different jurisdictions. Given multiple versions of the same contract, we selected the most recent version available on-line to European customers. The mark-up was done in XML by three annotators, which jointly worked for the formulation of the annotation guidelines. The whole annotation process included several revisions, where some corrections were also suggested by an analysis of the false positives and false negatives retrieved by the initial machine learning prototypes. Due to the large interaction among the annotators during this process, in order to assess inter-annotation agreement, a further test set consisting of 10 additional contracts was tagged, following the final version of the guidelines. We made the whole annotated corpus as well as the annotation guidelines available to the community, in an effort to encourage further research on this topic.[4]

3.1 Annotation process

In analyzing the Terms of Service of the selected on-line platforms, we identified eight different categories of unfair clauses, as described in Sect. 2. For each type of clause we defined a corresponding XML tag, as shown in Table 1.
Notice that not necessarily all the documents contain all clause categories. For example, Twitter provides two different ToS, the first one for US and non-US residents and the second one for EU residents. The tagged version is the version applicable in the EU and it does not contain any choice of law, arbitration or jurisdiction clauses.
We assumed that each type of clause could be classified as either clearly fair, or potentially unfair, or clearly unfair. In order to mark 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. Nested tags were used to annotate text segments relevant to more than one type of clause. With clauses covering multiple paragraphs, we chose to tag each paragraph separately, possibly with different degrees of (un)fairness.

[3] In particular, we selected the ToS offered by: 9gag.com, Academia.edu, Airbnb, Amazon, Atlas Solutions, Betterpoints, Booking.com, Crowdtangle, Deliveroo, Dropbox, Duolingo, eBay, Endomondo, Evernote, Facebook, Fitbit, Google, Headspace, Instagram, Linden Lab, LinkedIn, Masquerade, Microsoft, Moves-app, musically, Netflix, Nintendo, Oculus, Onavo, Pokemon GO, Rovio, Skype, Skyscanner, Snapchat, Spotify, Supercell, SyncMe, Tinder, TripAdvisor, TrueCaller, Twitter, Uber, Viber, Vimeo, Vivino, WhatsApp, World of Warcraft, Yahoo, YouTube and Zynga.
[4] http://claudette.eui.eu/ToS.zip.

Table 1 Categories of clause unfairness, with the corresponding symbol used for tagging

Type of clause	Symbol
Arbitration	<a>
Unilateral change	<ch>
Content removal	<cr>
Jurisdiction	<j>
Choice of law	<law>
Limitation of liability	<ltd>
Unilateral termination	<ter>
Contract by using	<use>

Jurisdiction This type of clause stipulates what courts will have the competence to adjudicate disputes under the contract. Jurisdiction clauses giving consumers a right to bring disputes in their place of residence were marked as clearly fair, whereas clauses stating that any judicial proceeding takes a residence away (i.e. in a different city, different country) were marked as clearly unfair. This assessment is grounded in ECJ's case law, see for example *Oceano* case number C-240/98. An example of jurisdiction clauses is the following one, taken from the Dropbox terms of service:

<j3>You and Dropbox agree that any judicial proceeding to resolve claims relating to these Terms or the Services will be brought in the federal or state courts of San Francisco County, California, subject to the mandatory arbitration provisions below. Both you and Dropbox consent to venue and personal jurisdiction in such courts.</j3>

<j1>If you reside in a country (for example, European Union member states) with laws that give consumers the right to bring disputes in their local courts, this paragraph doesn't affect those requirements.</j1>

The second clause introduces an exception to the general rule stated in the first clause, thus we marked the first one as clearly unfair and the second as clearly fair.

Choice of law This clause specifies what law will govern the contract, meaning also what law will be applied in potential adjudication of a dispute arising under the contract. Clauses defining the applicable law as the law of the consumer's country of residence were marked as clearly fair, as reported in the following examples, taken from the Microsoft services agreements:

<law1>If you live in (or, if a business, your principal place of business is in) the United States, the laws of the state where you live govern all

claims, regardless of conflict of laws principles, except that the Federal Arbitration Act governs all provisions relating to arbitration.</law1>

<law1>If you acquired the application in the United States or Canada, the laws of the state or province where you live (or, if a business, where your principal place of business is located) govern the interpretation of these terms, claims for breach of them, and all other claims (including consumer protection, unfair competition, and tort claims), regardless of conflict of laws principles.</law1>

<law1>Outside the United States and Canada. If you acquired the application in any other country, the laws of that country apply.</law1>

In every other case, the choice of law clause was considered as potentially unfair. This is because the evaluation of the choice of law clause needs to take into account several other conditions besides those specified the clause itself (for example, level of protection offered by the chosen law). Consider the following example, taken from the Facebook terms of service:

<law2>The laws of the State of California will govern this Statement, as well as any claim that might arise between you and us, without regard to conflict of law provisions</law2>

Limitation of liability This clause stipulates that the duty to pay damages is limited or excluded, for certain kinds of losses and under certain conditions. Clauses that explicitly affirm non-excludable providers' liabilities were marked as clearly fair. For example, consider the example below, taken from World of Warcraft terms of use:

<ltd1>Blizzard Entertainment is liable in accordance with statutory law (i) in case of intentional breach, (ii) in case of gross negligence, (iii) for damages arising as result of any injury to life, limb or health or (iv) under any applicable product liability act.</ltd1>

Clauses that reduce, limit, or exclude the liability of the service provider were marked as potentially unfair when concerning broad categories of losses or causes of them, such as any harm to the computer system because of malware or loss of data or the suspension, modification, discontinuance or lack of the availability of the service. Also those liability limitation clauses containing a blanket phrase like "to the fullest extent permissible by law", where considered potentially unfair. The following example is taken from 9gag terms of service:

<ltd2>You agree that neither 9GAG, Inc nor the Site will be liable in any event to you or any other party for any suspension, modification, discontinuance or lack of availability of the Site, the service, your Subscriber Content or other Content.</ ltd2>

Clause meant to reduce, limit, or exclude the liability of the service provider for physical injuries, intentional damages as well as in case of gross negligence were marked as clearly unfair (based on the Annex to the Directive) as showed in the example below, taken from the Rovio license agreement:

<ltd3>In no event will Rovio, Rovio's affiliates, Rovio's licensors or channel partners be liable for special, incidental or consequential damages resulting from possession, access, use or malfunction of the Rovio services, including but not limited to, damages to property, loss of goodwill, computer failure or malfunction and, to the extent permitted by law, damages for personal injuries, property damage, lost profits or punitive damages from any causes of action arising out of or related to this EULA or the software, whether arising in tort (including negligence), contract, strict liability or otherwise and whether or not Rovio, Rovio's licensors or channel partners have been advised of the possibility of such damages.<ltd3>

Unilateral change This clause specifies the conditions under which the service provider could amend and modify the terms of service and/or the service itself. Such clauses were always considered as potentially unfair. This is because the ECJ has not yet issued a judgment in this regard, though the Annex to the Directive contains several examples supporting such a qualification. Consider the following examples from the Twitter terms of service:

<ch2>As such, the Services may change from time to time, at our discretion.</ch2>

<ch2>We also retain the right to create limits on use and storage at our sole discretion at any time.</ch2>

<ch2>We may revise these Terms from time to time. The changes will not be retroactive, and the most current version of the Terms, which will always be at twitter.com/tos, will govern our relationship with you.</ch2>

Unilateral termination This clause gives provider the right to suspend and/ or terminate the service and/or the contract, and sometimes details the circumstances under which the provider claims to have a right to do so. Unilateral termination clauses that specify reasons for termination were marked as potentially unfair. whereas clauses stipulating that the service provider may suspend or terminate the service at any time for any or no reasons and/or without notice were marked as clearly unfair. That is the case in the three following examples, taken from the Dropbox and Academia terms of use, respectively:

<ter2>We reserve the right to suspend or terminate your access to the Services with notice to you if: (a) you're in breach of these Terms, (b) you're using the

Services in a manner that would cause a real risk of harm or loss to us or other users, or (c) you don't have a Paid Account and haven't accessed our Services for 12 consecutive months.</ ter2>

<ter3>Academia.edu reserves the right, at its sole discretion, to discontinue or terminate the Site and Services and to terminate these Terms, at any time and without prior notice.</ter3>

Contract by using This clause stipulates that the consumer is bound by the terms of use of a specific service, simply by using the service, without even being required to mark that he or she has read and accepted them. We always marked such clauses as potentially unfair. The reason for this choice is that a good argument can be offered for these clauses to be unfair, because they originate an imbalance in rights and duties of the parties, but this argument has no decisive authoritative backing yet, since the ECJ has never assessed a clause of this type. Consider an example taken from the Spotify terms and conditions of use:

<use2>By signing up or otherwise using the Spotify service, websites, and software applications (together, the "Spotify Service" or "Service"), or accessing any content or material that is made available by Spotify through the Service (the "Content") you are entering into a binding contract with the Spotify entity indicated at the bottom of this document.</use2>

Content removal This gives the provider a right to modify/delete user's content, including in-app purchases, and sometimes specifies the conditions under which the service provider may do so. As in the case of unilateral termination, clauses that indicate conditions for content removal were marked as potentially unfair, whereas clauses stipulating that the service provider may remove content in his full discretion, and/or at any time for any or no reasons and/or without notice nor possibility to retrieve the content were marked as clearly unfair. For instance, consider the following examples, taken from Facebook's and Spotify's terms of use:

<cr2>If you select a username or similar identifier for your account or Page, we reserve the right to remove or reclaim it if we believe it is appropriate (such as when a trademark owner complains about a username that does not closely relate to a user's actual name).</cr2>

<cr2>We can remove any content or information you post on Facebook if we believe that it violates this Statement or our policies.</cr2>

<cr3>In all cases, Spotify reserves the right to remove or disable access to any User Content for any

or no reason, including but not limited to, User Content that, in Spotify's sole discretion, violates the Agreements. Spotify may take these actions without prior notification to you or any third party.</ cr3>

Arbitration This clause requires or allows the parties to resolve their disputes through an arbitration process, before the case could go to court. It is therefore considered a kind of forum selection clause. However, such a clause may or may not specify that arbitration should occur within a specific jurisdiction. Clauses stipulating that the arbitration should (1) take place in a state other than the state of consumer's residence and/or (2) be based not on law but on arbiter's discretion were marked as clearly unfair. As an illustration, consider the following clause of the Rovio terms of use:

<j1><a3>Any dispute, controversy or claim arising out of or relating to this EULA or the breach, termination or validity thereof shall be finally settled at Rovio's discretion (i) at your domicile's competent courts; or (ii) by arbitration in accordance with the Rules for Expedited Arbitration of the Arbitration Institute of the Finland Chamber of Commerce. The arbitration shall be conducted in Helsinki, Finland, in the English language.</a3></j1>

Notice that the clause above concerns both jurisdiction and arbitration (thus the use of nested tags). Clauses defining arbitration as fully optional would have to be marked as clearly fair. However, our corpus does not contain any example of fully optional arbitration clause. Therefore, all arbitration clauses were marked as potentially unfair. An example is the following segment of Amazon's terms of service:

<a2>Any dispute or claim relating in any way to your use of any Amazon Service, or to any products or services sold or distributed by Amazon or through Amazon.com will be resolved by binding arbitration, rather than in court, except that you may assert claims in small claims court if your claims qualify. The Federal Arbitration Act and federal arbitration law apply to this agreement.</a2>

Table 2 Corpus statistics

Type of clause	# clauses	# documents
Arbitration	44	28
Unilateral change	188	49
Content removal	118	45
Jurisdiction	68	40
Choice of law	70	47
Limitation of liability	296	49
Unilateral termination	236	48
Contract by using	117	48

For each category of clause unfairness, we report the overal number of clauses and the number of documents they appear in

3.2 Corpus statistics

The corpus contains 12,011 sentences,[5] 8.6% of which (1,032 sentences) were labeled as positive, thus containing a potentially unfair clause. The distribution of the different categories across the 50 documents is reported in Table 2. Arbitration clauses are most uncommon, and are found in 28 documents only. All other categories appear in at least 40 out of 50 documents. Limitation of liability and unilateral termination together represent more than half of all potentially unfair clauses. The percentage of potentially unfair clauses in each document is quite heterogeneous, ranging from 3.3% (Microsoft) up to 16.2% (TrueCaller).

3.3 Additional test set

We produced an additional test set consisting of 10 more annotated contracts.[6] Such documents were independently tagged by two distinct annotators who had carefully studied the guidelines. In order to quantitatively measure the inter-annotation agreement, for this test set we computed the standard Cohen's metric (Cohen 1968), which resulted to be 0.871, a value that is typically considered as an "almost perfect agreement" (Landis and Koch 1977). This second test set was used for a further evaluation of the deployed system.

[5] Segmentation into sentences was made using the Stanford CoreNLP suite (see Sect. 5).
[6] In particular, we selected the ToS offered by: Alibaba, Badoo, Goodreads, Groupon, Mozilla, Ryanair, Shazam, Slack, Zalando UK, eDreams.

4 Machine learning methodology

In this section we briefly describe the representation and learning methods used in our study.
The study focuses two different tasks: a detection task, aimed at predicting whether a given sentence contains a (potentially) unfair clause, and a classification task, aimed at predicting the category an unfair clause belongs to, which indeed could be a valuable piece of information to a potential user. Results on the two tasks are presented in Sect. 5.

4.1 Learning algorithms

We address the problem of detecting potentially unfair contract clauses as a sentence classification task. Such a task could be tackled by treating sentences independently of one another (*sentence-wide* classification). This is the most standard and classic approach in machine learning, traditionally addressed by methods such as Support Vector Machines or Artificial Neural Networks, which include recent deep learning approaches (Kim 2014).
Alternatively, one could take into account the structure of the document, in particular the *sequence* of sentences, so as to perform a *collective* classification, as it has been done in cognate sentence classification tasks (Habernal and Gurevych 2017). The potential advantage of such an approach becomes apparent if we observe that unfair clauses often span across consecutive sentences in a document.

In sentence-wide classification, the problem can be formalized as follows. Given a sentence, the goal is to classify it as *positive* if it contains a potentially unfair clause, or *negative* otherwise. Within this setting, a machine learning classifier is trained with a data set $D = \{(x_i, y_i)\}_{i=1}^{N}$, consisting of a collection of N pairs, where x_i encodes some representation of a sentence, and y_i is its corresponding (positive or negative) class.

In collective classification, the data set consists of a collection of M *documents*, represented as sequences of sentences:

$$\mathcal{D} = \left\{ d_j = \left\{ \left(x_1^j, y_1^j \right), \ldots, \left(x_{k_j}^j, y_{k_j}^j \right) \right\} \right\}_{j=1}^{M},$$

where the jth document contains k_j sentences.

Different machine learning systems can be developed for each classification setup, according to the learning framework and to the features employed to represent each sentence. As for the learning methodology, for sentence-wide classification in this paper we compare Support Vector Machines (SVMs) (Joachims 1998) with some recent deep learning architectures, namely Convolutional Neural Networks (CNNs) (Kim 2014) and Long-Short Term Memory Networks (LSTMs) (Graves and Schmidhuber 2005). For collective classification, we rely on structured Support Vector Machines, and in particular on SVM-HMMs, which combine SVMs with Hidden Markov Models (Tsochantaridis et al. 2005), by jointly assigning a label to each element in a given sequence (in our case, to each sentence in the considered document).

4.2 Sentence representation

As for the features represented to encode sentences, in an effort to make our method as general as possible, we decided to opt for traditional features for text categorization, excluding other, possibly more sophisticated, handcrafted features.

One of the most classic, yet still widely used, set of features for text categorization, is the well-known *bag-of-words* (BoW) model. In such a model, one feature is associated with each word in the vocabulary: the value of such a feature is either zero, if the word does not appear in the sentence, or other than zero, if it does. Such a value is usually computed as the TF-IDF score, that is the number of occurrences of the word in the sentence (Term Frequency, TF) multiplied by a term that amplifies the weight of infrequent words (Inverse Document Frequency, IDF) (Sebastiani 2002).

The BoW model can be extended to consider also *n*-grams, i.e., consecutive word combinations, rather than simple words, so as to exploit, at least locally, the ordering of words in the sentences. Grammatical information can be included as well, by constructing a bag of part-of-speech tags, i.e., word categories such as nouns, verbs, etc. (Leopold and Kindermann 2002). Despite their simplicity, BoW features are very informative, as they encode the lexical information of a sentence, and thus represent a challenging baseline in those cases where the presence of some keywords and phrases is highly discriminative for the categorization of sentences.

A second approach we consider for the representation of a sentence exploits a constituency parse tree, which naturally encodes the *structure* of the sentence (see Fig. 1) by describing the grammatical relations between sentence portions through a tree. Similarity between tree structures can be exploited using *tree kernels* (Moschitti 2006) (TK). A TK consists of a *similarity measure* between two trees, which takes into account the number of common substructures, known as *fragments*. Different definitions of fragments induce different TK functions. In our study we use the SubSet Tree Kernel (SSTK) (Collins and Duffy 2002) which counts as fragments those subtrees of the constituency parse tree terminating either at the leaves or at the level of non-terminal symbols. SSTK have been shown to outperform other TK functions in several argumentation mining sub-tasks (Lippi and Torroni 2016b).

A third approach for sentence representation is based on *word embeddings* (Mikolov et al. 2013), a popular technique that has been recently developed in the context of neural language models and deep learning applications. Neural networks such as CNNs and LSTMs can handle textual input, by converting it into a sequence of identifiers, one for each different word. The neural network then directly learns a vector representation or "embedding" of words and sentences.

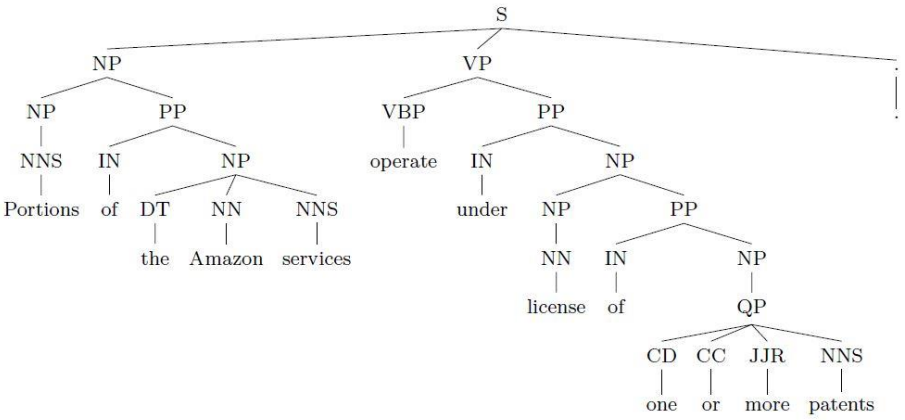

Fig. 1 An example of a constituency parse tree for a sentence in our corpus

5 Experimental results

We evaluated and compared several machine learning systems on the data set presented in Sect. 3. Each document was segmented into sentences, tokenized and parsed with the Stanford CoreNLP tool[7]. We discarded sentences and text fragments with less than 5 words. We thus obtained 9414 sentences, 11.0% of which, amounting to 1032 sentences, were labeled as positive, thus containing a potentially unfair clause.

We run experiments following the *leave-one-document-out* (LOO) procedure, in which each document in the corpus, in turn, is used as test set, leaving the remaining documents for training set (4/5) and validation set (1/5) for model selection. This is a standard procedure in machine learning, as it allows to assess the generalization capabilities of our system. The adoption of such a procedure, together with the high inter-annotation agreement achieved during the creation of the corpus, contribute to strengthen the validity of our experimental results.

To quantitatively evaluate the different tested classifiers, we computed precision (P) as the fraction of positive predictions, which are actually labeled as positive, recall (R) as the fraction of positive examples that are correctly detected, and finally F_1 as the harmonic mean between precision and recall ($F_1 = \frac{2PR}{P+R}$). These performance measurements were aggregated using the macro-average over documents (Sebastiani 2002). In principle, if the goal was to obtain a *complete* set of potentially unlawful clauses, a recall under 100% would require the user to scan the whole document. However, if the price of a 100% recall was a very low precision, the tool would clearly lose its purpose. By contrast, if the goal was to obtain a *correct* (though not necessarily exhaustive) set of potentially unlawful clauses, then one should prefer a high precision. For these reasons, we optimized the machine learning hyper-parameters based on the F_1 score, which is a customary trade-off between R and P. For neural architectures, we tested several configurations and chose the network achieving the best results on the validation set.

5.1 Detection of potentially unfair clauses

For the first task (potentially unfair clause detection) we compared several systems. The problem is formulated as a binary classification task, where the positive class is either the union of all potentially unfair sentences, or the set of potentially unfair clauses of a single category, as described below. We considered the following systems:

C1: A single SVM exploiting BoW (unigrams and bigrams for words and part-of-speech tags);
C2: A combination of eight SVMs (same features as above), each considering a single unfairness category as the positive class, whereby a sentence is predicted as potentially unfair if at least one of the SVMs predicts it as such;
C3: A single SVM exploiting TK for sentence representation; C4: A CNN trained from plain word sequences;

[7] [7] https://stanfordnlp.github.io/CoreNLP/..

C5: An LSTM trained from plain word sequences;
C6: An SVM-HMM performing collective classification of sentences in a document (word unigrams, bigrams, and trigrams);
C7: A combination of eight SVM-HMMs, each performing collective classification of sentences in a document on a single unfairness category as the positive class (same features as C6);
C8: An ensemble method, which combines the output of C1, C2, C3, C6 and C7 with a voting procedure (sentence predictive as positive if at least 3 systems out of 5 classify it as such).

As a reference for the complexity of the task, we also report the performance of the following baselines: a *random* classifier, which predicts potentially unfair clauses at random,[8] and an *always positive* baseline, which classifies every sentence as potentially unfair. For all the classifiers, the validation set was used to select the best hyper-parameters. For all SVMs we used a linear kernel, thus only optimizing the *C* hyper-parameter, which is responsible for regularization and thus for the generalization capabilities of the classifier. For SVM-HMM we used an order of dependencies equal to 2 and 1 for transitions and emissions, respectively; different from SVMs, we also used trigrams besides unigrams and bigrams, as they slightly increased performance. For CNNs, we considered one layer with 64 filters of size equal to 3, followed by two fully connected layers with 32 and 16 neurons, respectively. We applied dropout equal to 0.5, batch size equal to 16. An embedding of size 64 was learned after the input layer. For LSTMs, we considered a 2-layer network with 64 and 32 cells, respectively, with 0.25 dropout and mini-batch size equal to 16. An embedding of size 32 was learned after the input layer. Both for CNNs and LSTMs, no improvement was observed if using pre-trained word embeddings.

Table 3 Results on leave-one-document-out procedure

Classifier	Method	*P*	*R*	F_1
C1	SVM—single model	0.729	**0.830**	0.769
C2	SVM—combined model	0.798	0.782	0.781
C3	Tree kernels	0.777	0.718	0.739
C4	Convolutional neural networks	0.729	0.739	0.722
C5	Long short-term memory networks	0.696	0.723	0.698
C6	SVM-HMM—single model	0.759	0.778	0.758
C7	SVM-HMM—combined model	**0.859**	0.687	0.757
C8	Ensemble (C1+C2+C3+C6+C7)	0.826	0.797	**0.805**
	Random baseline	0.125	0.125	0.125
	Always positive baseline	0.123	1.000	0.217

Best results are highlighted in bold

[8] Sampling takes into account the class distribution in the training set.

Table 3 shows the results achieved by each of these variants. If we exclude the ensemble approach, the best classifier in terms of F_1 results to be C2, that is the system combining one different SVM trained for each unfairness category, with a precision above 80%, and a recall of 78%. The structured SVMs exploiting the sequentiality of the sentences achieve slightly lower results, yet very interestingly the results of the sentence-wise and document-wise approaches are different across different documents. Moreover, the worse performance associated with TK suggests that the syntactic structure of the sentence is less informative than the lexical information captured by *n*-grams. This makes the task of detecting unfair clauses different from other text retrieval problems in the legal domain, such as, for example, the detection of claims and arguments (Lippi and Torroni 2016a). As for CNNs and LSTMs, the slightly worse performance with respect to the other approaches could also be ascribed to the limited size of the training set. Nevertheless, we intend to investigate more sophisticated deep learning approaches in the future.

All these observations led us to the implementation of an ensemble method (C8), combining the five best performing approaches. This system achieves an F_1 of around 81%, outperforming all competitors. Such a result is particularly interesting, because it confirms that the different systems capture complementary information for the detection of potentially unfair clauses. The ensemble method correctly detects around 80% of the potentially unfair clauses in each category, ranging from a minimum 72.7% in the case of arbitration clauses, up to 89.7%, as in the case of jurisdiction clauses.

In order to better understand which *n*-grams contribute the most to the discrimination between fair and potentially unfair clauses, we computed the frequencies of bigrams in both positive and negative support vectors of classifier C1, and we looked for those with the largest discrepancy in appearing in the positive class rather than in the negative one. Some of the most salient bigrams, according to such a ranking, were: *for any*, *the right*, *these terms*, *any time*, *at any*, *right to*, *reserves the*, *we may*, *liable for*, *terminate your*, *sole discretion*, *the services*. This analysis confirms that the discriminative lexicon is quite general and widespread both across the different unfairness categories and the different types of services we considered.

As a further evaluation of our approach, we used the additional test set of 10 documents described in Sect. 3.3. We obtained a macro-average precision, recall and F_1 of the ensemble system equal to 0.782, 0.708 and 0.736, respectively.

Table 4 Micro-averaged precision, recall and F_1 of abusive clauses for each tag category

Tag	Precision	Recall	F_1
Arbitration	0.832	0.814	0.823
Unilateral change	0.832	0.814	0.823
Content removal	0.713	0.780	0.745
Jurisdiction	1.000	0.941	0.970
Choice of law	0.984	0.886	0.932
Limitation of liability	0.961	0.905	0.932
Unilateral termination	0.786	0.932	0.853
Contract by using	0.949	0.957	0.953

5.2 Categorization of potentially unfair clauses

The second task we considered is unfairness categorization, for which we employed eight SVM classifiers, each trained to discriminate between potentially unfair clauses of one category with respect to all the other categories. It is worthwhile remarking that this task differs from that addressed by the previously introduced classifiers, since in this case the classifiers are trained on potentially unfair clauses only. Moreover, this task is a *multi-label* classification and not a *multiclass* task, because each sentence can potentially belong to several unfairness categories. In Table 4 we report the precision, recall, and F_1 of such classifiers, one for each separate tag category, micro-averaged on the whole dataset. The results show that discriminating amongst the different categories is a simpler task, since the F_1 is larger than 74% for all tags, and is above 93% in four cases (jurisdiction, choice of law, limitation of liability, and contract by using).

5.3 Error analysis

In an effort to understand which kinds of sentences are harder to classify, we run a qualitative analysis considering the false positives and false negatives produced by the system. A significant number of errors concerns sentences about third parties: around 10% of both false positives and false negatives contain in fact the keyword "third party". Clearly, these are challenging sentences, because the treatment of data collected by third parties can or cannot, in principle, be compliant with the law. Consider for instance the following clauses—respectively, a false positive and a false negative:

> You understand and agree that Spotify does not endorse and is not responsible or liable for the behavior, features, or content of any third party application or for any transaction you may enter into with the provider of any such third party applications.

> Skype may, without prior notice, assign these terms or any rights or obligations contained in them to any third party.

Another set of sentences that contributes to a significant number of errors has to do with the responsibility of damages. In particular, this group of sentences produces a quite large set of false positives (14%) and a much smaller set of false negatives (5%). This means that CLAUDETTE tends to over-predict potentially unlawful clauses, when the sentence refers to the responsibility in case of damages. One clear example of such a false positive is given by the following case:

> Crowdtangle will not be responsible for any loss or damages resulting from your failure to comply with this obligation or otherwise any unauthorized use of your account.

A large portion of false negatives (over 18%) concerns practices related to the term "content" from different perspectives, such as content removal, liability for content publication, responsibility for content integrity, correctness and appropriateness. The following is an excerpt from Deliveroo:

```
Generally, we do not moderate any interactive service
we provide although we may remove content in
contravention of these terms of use as set out in
Section 6.
```

One possible direction for future research is to consider a specific rule-based module as a post-processing phase of CLAUDETTE, to handle some of the aforementioned error categories.

6 The CLAUDETTE web server

The proposed approach was implemented and developed as a web server, reachable at the address http://claudette.eui.eu/demo, so as to produce a prototype system that users can easily access and test.

As shown in Fig. 2, the interface is easy to use. A user only needs to paste the text to be analyzed and push a button. The system will then produce an output file that highlights the sentences predicted to contain a potentially unfair clause. The output will also indicate the predicated category the unfair clause belongs to, as illustrated in Fig. 3. The output of the system can be obtained in several formats including HTML, XML, JSON, and plain text.

For this online service, for the detection stage we implemented only one system (namely, classifier C2) rather than the ensemble method, because it resulted to be a much more efficient solution in terms of running time, despite producing a slightly lower performance accuracy.

Fig. 2 The interface of the CLAUDETTE web server, consisting of a box where a user can copy–paste the text of a terms of service

7 Related work

The use of artificial intelligence, machine learning and natural language processing techniques in the analysis and classification of legal documents is gaining a growing interest (Ashley 2017). Among others, Moens et al. (2007) proposed a pipeline of steps for the extraction of arguments from legal documents, exploiting supervised classifiers and context-free grammars, whereas Biagioli et al. (2005) proposed to employ multi-class SVM for the identification of significant text portions in normative texts. Recent approaches have focused on the detection of claims (Lippi et al. 2018) and of cited facts and principles in legal judgments (Shulayeva et al. 2017), as well as on the prediction of judicial decisions (Aletras et al. 2016) and legal compliance assessment (Bartolini et al. 2016; Robaldo and Sun 2017). A case study regarding the construction of legal arguments in the legal determinations of vaccine / injury compensation compliance using natural language tools was given by Ashley and Walker (2013). Finally, privacy policies represent another closely related and increasingly popular application domain, where machine learning approaches have proven effective, as discussed by Fabian et al. (2017) and references therein, as well as by Harkous et al. (2018).
Typically, applications in this domain address the problem of categorizing text portions in privacy policies, with the aim of summarizing or extracting relevant information from such documents, to improve readability for the end-user. Differently from our approach, the legal task of detecting unfairness is usually not taken into account.

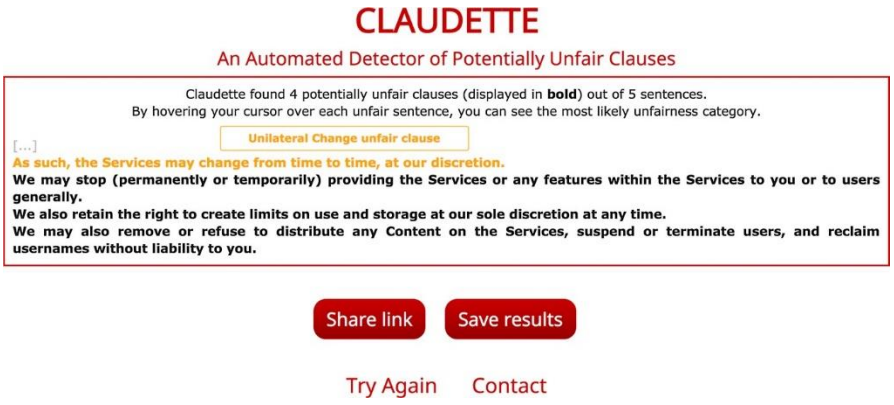

Fig. 3 Results of a query to the CLAUDETTE web server. Hovering over a detected clause with the pointer provides an indication of the type of potentially unfair clause. In this example the detected clauses are predicted to be of types *unilateral change*, *unilateral termination*, and *content removal*, and the cursor was left hovering over the first potentially unfair clause

8 Conclusions

Our study investigates the use of machine learning and natural language methods for the automated detection of potentially unfair clauses in online contracts. We addressed two tasks: clause detection and clause type classification. For clause detection, our results are very encouraging: using a relatively small training set we could automatically detect over 80% clauses, with an 80% precision. The categorization task turned out to be simpler. Given that most unfair clauses are currently hidden within long and hardly readable ToS, the recall and precision offered by our approach may already be significant enough to enable useful applications.

It is interesting to notice the comparatively better performance of the BoW approach with respect to other more sophisticated approaches. That is in agreement with the surveyed literature, where classic lexical approaches such as BoW still represent a crucial ingredient of automated systems. It is also worth remarking the best performance yielded by an ensemble method, indicating that different machine learning approaches are capable of capturing diverse characteristics of potentially unfair clauses.

This study was motivated by a long-term goal concerned with the pursuit of effective consumer protection by way of AI-based consumer-empowering tools. The CLAUDETTE system represents our first step in this direction, and it shows that machine learning tools can help the civil society in monitoring on-line terms of services. To further that goal, we are collaborating with consumer organizations towards the development of a more user-friendly version of our system to be made available on-line. We are also working on new developments and extensions. In particular, we are investigating methods for exploiting contextual information, since the fairness of clauses might very well depend on the context. For example, a potentially unfair jurisdiction clause might actually be fair according to EU regulation if is followed by a paragraph stipulating relevant exceptions according to the user's country of residence. Another challenging line of research we are pursuing is the adaptation of the methodology used for CLAUDETTE in order to enable the automated analysis of privacy policies: an important area of consumer protection which gained recent media focus due to its enormous implications for individuals and for the society at large.

Acknowledgements Funding was obtained from European University Institute by author Hans-Wolfgang Micklitz (CLAUDETTE Project).

References

Aletras N, Tsarapatsanis D, Preoiuc-Pietro D, Lampos V (2016) Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. PeerJ Comput Sci 2:e93

Ashley K (2017) Artificial intelligence and legal analytics: new tools for law practice in the digital age. Cambridge University Press, Cambridge

Ashley KD, Walker VR (2013) Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In: Francesconi E, Verheij B (eds) ICAIL 2012, Rome, Italy, ACM, pp 176–180. https://doi.org/10.1145/2514601.2514622. http://dl.acm.org/citation.cfm?id=2514622

Bakos Y, Marotta-Wurgler F, Trossen DR (2014) Does anyone read the fine print? Consumer attention to standard-form contracts. J Legal Stud 43(1):1–35

Bartolini C, Giurgiu A, Lenzini G, Robaldo L (2016) Towards legal compliance by correlating standards and laws with a semi-automated methodology. In: BNCAI, Communications in computer and information science, vol 765. Springer, pp 47–62

Biagioli C, Francesconi E, Passerini A, Montemagni S, Soria C (2005) Automatic semantics extraction in law documents. In: Proceedings of ICAIL, ACM, pp 133–140

Cohen J (1968) Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull 70(4):213

Collins M, Duffy N (2002) New ranking algorithms for parsing and tagging: kernels over discrete structures, and the voted perceptron. In: Proceedings of the 40th annual meeting of the ACL, ACL, pp 263–270

Department of Commerce (2010) Commercial data privacy and innovation in the internet economy: a dynamic policy framework. Technical report, Department of Commerce Internet Policy Task Force. https://www.ntia.doc.gov/files/ntia/publications/iptf_privacy_greenpaper_12162010.pdf

Fabian B, Ermakova T, Lentz T (2017) Large-scale readability analysis of privacy policies. In: Proceedings of the international conference on web intelligence, ACM, pp 18–25

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural Netw 18(5):602–610

Habernal I, Gurevych I (2017) Argumentation mining in user-generated web discourse. Comput Linguist 43(1):125–179

Harkous H, Fawaz K, Lebret R, Schaub F, Shin KG, Aberer K (2018) Polisis: automated analysis and presentation of privacy policies using deep learning. arXiv:180202561

Joachims T (1998) Text categorization with support vector machines: Learning with many relevant features. In: ECML, vol 98, pp 137–142

Kim Y (2014) Convolutional neural networks for sentence classification. In: Moschitti A, Pang B, Daelemans W (eds) Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, October 25–29, 2014, Doha, Qatar, A meeting of SIGDAT, a special interest group of the ACL, ACL, pp 1746–1751

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174

Leopold E, Kindermann J (2002) Text categorization with support vector machines. How to represent texts in input space? Mach Learn 46(1–3):423–444

Lippi M, Torroni P (2016a) Argumentation mining: state of the art and emerging trends. ACM Trans Internet Technol 16(2):10:1–10:25

Lippi M, Torroni P (2016b) Margot: a web server for argumentation mining. Expert Syst Appl 65(C):292–303. https://doi.org/10.1016/j.eswa.2016.08.050

Lippi M, Palka P, Contissa G, Lagioia F, Micklitz H, Panagis Y, Sartor G, Torroni P (2017) Automated detection of unfair clauses in online consumer contracts. In: Wyner AZ, Casini G (eds) Legal knowledge and information systems—JURIX 2017: the thirtieth annual conference, vol. 302, Luxembourg, 13–15 December 2017, IOS Press, Frontiers in Artificial Intelligence and Applications, pp 145–154

Lippi M, Lagioia F, Contissa G, Sartor G, Torroni P (2018) Claim detection in judgments of the EU Court of Justice. In: Artificial intelligence and the complexity of legal systems, VI international workshop (AICOL), selected revised papers. Lecture notes in artificial intelligence, Springer, forthcoming

Loos M, Luzak J (2016) Wanted: a bigger stick. On unfair terms in consumer contracts with online service providers. J Consum Policy 39(1):63–90

McDonald A, Cranor L (2008) The cost of reading privacy policies. I/S J Law Policy Inf Soc 4(3):543–568

Micklitz HW, Reich N (2014) The court and sleeping beauty: the revival of the unfair contract terms directive (UCTD). Common Market Law Rev 51(3):771-808

Micklitz HW, Pałka P, Panagis Y (2017) The empire strikes back: digital control of unfair terms of online services. J Consum Policy 40(3):367–388

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arxiv: 1301.3781
Moens MF, Boiy E, Palau RM, Reed C (2007) Automatic detection of arguments in legal texts. In: Proceedings of the 11th international conference on artificial intelligence and law, ACM, pp 225–230

Moschitti A (2006) Efficient convolution kernels for dependency and constituent syntactic trees. In: Fürnkranz J, Scheffer T, Spiliopoulou M (eds) Machine learning: ECML 2006, LNCS, vol 4212. Springer, Berlin Heidelberg, pp 318–329

Nebbia P (2007) Unfair contract terms in European law: a study in comparative and EC law. Bloomsbury Publishing, London

Obar JA, Oeldorf-Hirsch A (2016) The biggest lie on the internet: ignoring the privacy policies and terms of service policies of social networking services. In: TPRC 44: the 44th research conference on communication, information and internet policy

Reich N, Micklitz HW, Rott P, Tonner K (2014) European consumer law. Intersentia, Cambridge

Robaldo L, Sun X (2017) Reified input/output logic: combining input/output logic and reification to represent norms coming from existing legislation. J Logic Comput 27(8):2471–2503

Schulte-Nölke H, Twigg-Flesner C, Ebers M (2008) EC consumer law compendium: the consumer acquis and its transposition in the member states. Walter de Gruyter, Berlin

Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47. https://doi.org/10.1145/505282.505283

Shulayeva O, Siddharthan A, Wyner A (2017) Recognizing cited facts and principles in legal judgements. Artif Intell Law 25(1):107–126

Tsochantaridis I, Joachims T, Hofmann T, Altun Y (2005) Large margin methods for structured and interdependent output variables. J Mach Learn Res 6:1453–1484.