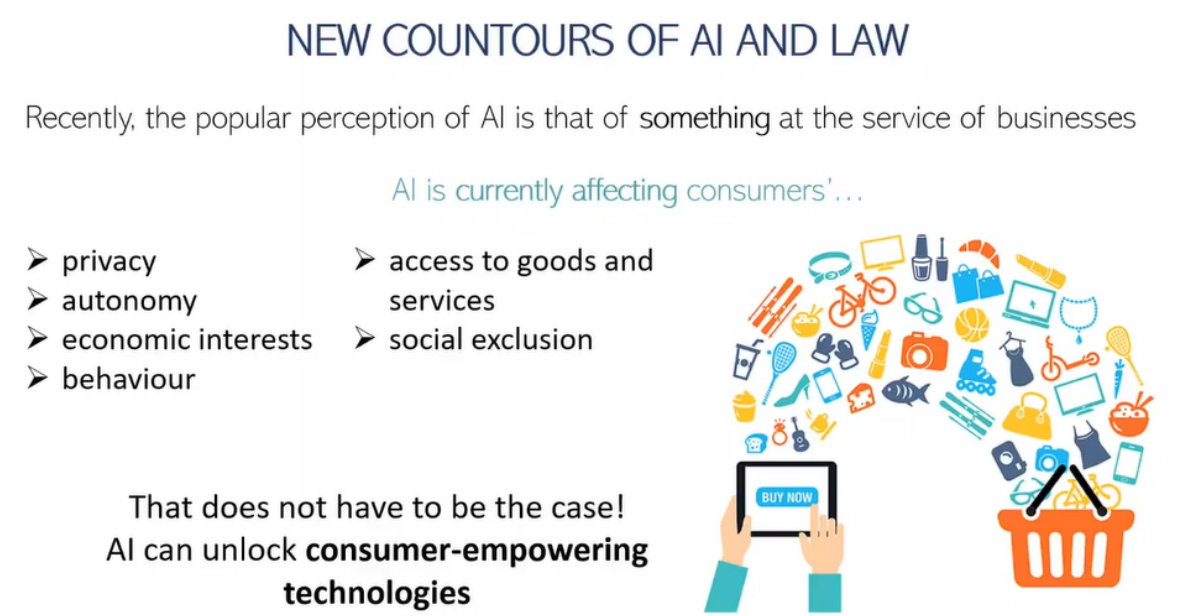
Lesson_11_Claudette_System

The Claudette System

Automation of personal data and consumer law enforcement using AI

This tool has been developed because we reveal a huge amount of personal data on the web and among them there are *sensitive* information



AI used by organizations who gather those information affect individuals, particularly the customers, in various ways (slide). This is the reason behind the developing of Claudette: its developers believe that AI can be used to unlock consumers empowering technology.

How to empower consumers?

- **Protection** against unwanted monitoring (GDPR)
- **Support** in detecting unfair use of AI
- **Control** commercial practice fairness

> *"An opposing exercise of power is the principal solvent of economic power, the basic defense against its exercise in economic affairs". Ken Galbraith*

In the AI era an effective countervailing power needs to be supported by AI.

CLAUDETTE.eui.eu

Automatically detect **potentially** unfair clauses in Terms of Services and Privacy Policies

- Consumers agree but don't read
- NGOs have competence to control but lack resources
- Business keeps using unlawful clauses

Claudette is a machine learning system based on supervised learning that could potentially detect **unfair clauses** in Terms of Services (ToS) and Privacy Policies (which nobody reads, let's be honest), which may be present even though regulations on their content exists. [NGO: non governmental organization].

Terms of Service (ToS): The Training Set

The ToS Corpus

WHERE DID WE START?

... 50 ToS (manually annotated)...

7,090 sentences, 787 of which (11.1%) were labeled as positive, thus containing a potentially unfair clause.

WHERE ARE WE NOW?

... 100 ToS (manually annotated)...

In the slide above we can see from which services they took the terms. To decide which terms should be included in the dataset they used a *geographical criterion*, the *time of service enstablishment*, the *number of user* and so on.

Part 1: Unfair Contract Terms Law and Practice

Directive 93/13 art 3.1:

A contractual term which has **not** been **individually negotiated** shall be regarded as **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.

Bottom-line: there are some types of clauses that traders are prohibited from using in the contracts.

To label these documents, they started from the **general definition of unfairness** contained in the *unfair contract terms directive.* In the slide above we can see one article from it.

8 unfairness categories
(Art. 3 of Directive 93/13)

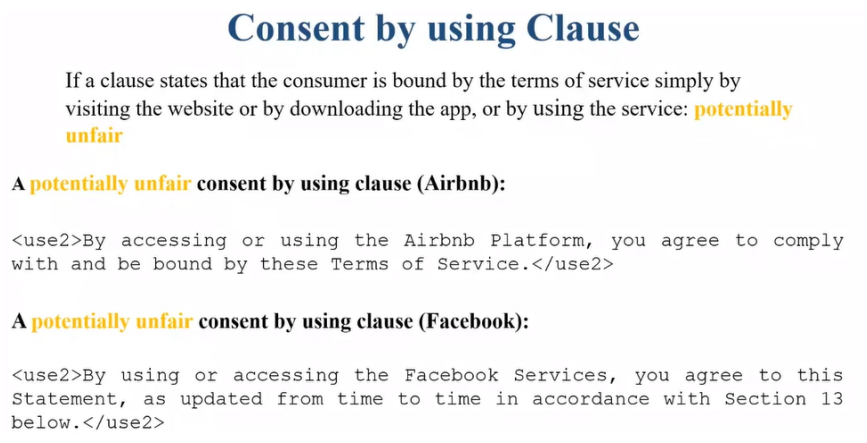
Type of clause	Symbol (xml tag)	# clauses (50 Tos)	#documents (50 Tos)
Arbitration	<a>	44	28
Unilateral change	<ch>	188	49
Content removal	<c>	118	45
Jurisdiction	<j>	68	40
Choice of law	<law>	70	47
Limitation of liability	<ltd>	296	49
Unilateral termination	<ter>	236	48
Contract by using	<use>	117	48

1) clearly fair; 2) potentially unfair; 3) clearly unfair

On that general definition, they defined the categories of clauses they wanted to look for and analyze.

- **Unilateral change**: clauses that state that the service provider can **unilaterally change** the contract without the user's consent.
- **Content Removal**: the service provider is allowed to remove unilaterally some content.
- **Jurisdiction Clauses:** related to the court, in the case the user wants to start a judicial proceeding.
- **Choice of law**: which law is applicable in case law (giurisprudenza).

To each clause type they associated a *xml* tag and to each tag they appended a grade of fairness. (e.g, j3 Jurisdiction clause *clearly unfair*)



Why the second clause is potentially unfair aside from the *consent by using* clause? Simply by visiting the site, not only we agree to the ToS contract but it can be modified in the future and we should check from time to time if there are any changes that affect us.

Jurisdiction Clause

Where a dispute will be adjudicated?

If giving consumers a right to bring disputes in their place of residence: **clearly fair**

If stating that any judicial proceeding takes a residence away (i.e. in a different city, different country): **clearly unfair**

A clearly unfair jurisdiction clause (Dropbox):

```
<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>
```

If the clause states that any judicial proceeding has to be held in a place different from the user place of residence, the clause has been considered *clearly unfair.* This is due to the fact that the user might decide to not start a proceeding because it would be too expansive to reach the location reported in the clause.

Limitation of Liability

For what actions/events the provider claims they will not be liable?

If stating that the provider may be liable: **clearly fair**

If stating that the provider will never be liable for any action taken by other people// damages incurred by the computer because of malware // When contains a blanket phrase like “to the fullest extent permissible by law”: **potentially unfair**

If stating that the provider will never be liable for physical injuries (health/life)// gross negligence// intentional damage: **clearly unfair**

Why the guideline on potential unfairness of this kind of clauses has been defined that way?
There might be some ambiguities: a clause might be fair if there are other clauses

that clarify it. The potentially unfair class covers all the cases where the clause could possibly create an unbalance between the signing parts.

Limitation of Liability

For what actions/events the provider claims they will not be liable?

A fair jurisdiction clause (World of Warcraft):

<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>

A potentially unfair jurisdiction clause (9gag):

<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>

Why the potentially unfair?

For example, the lack of availability of the website could cause damage to the consumer who might have need to access to some content or maybe has payed for that service. There are some situation that would justify the service interruption (force majeure), but it's not the case of this clause.

Limitation of Liability

For what actions/events the provider claims they will not be liable?

A clearly unfair jurisdiction clause (Rovio):

<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>

Rovio is never liable basically. lol.

An example from the Instagram Terms of Service

We reserve the right, in our sole discretion, to change these Terms of Use ("Updated Terms") from time to time.

Unless we make a change for legal or administrative reasons, we will provide reasonable advance notice before the Updated Terms become effective. You agree that we may notify you of the Updated Terms by posting them on the Service, and that your use of the Service after the effective date of the Updated Terms (or engaging in such other conduct as we may reasonably specify) constitutes your agreement to the Updated Terms.

What is wrong in this ToS?

- The first clause falls into the **unilateral change** category, as the company reserves the right to itself to change the ToS in *their sole discretion.*
- It is a *consent by using* clause because it states that if we keep using the service after the ToS change, we agree to it.

System Training

The Machine Learning Methodology

From a ML point of view, we modelled the problem as:

a **detection task**: does a sentence contain a potentially unfair clause? Positive (if p unfair), Negative (otherwise)

a **sentence classification task**: what is the category the unfair clause belongs to?

Approaches

- Bag of Words (BoW): build to leverage the lexical information in sentences
- Tree kernels: structure of sentences by describing the grammatical relations between sentence through a tree
- Convolutional Neural Networks, SVM, etc.

Experiments

Leave-One-Out procedure: each document in turn, is used as test set, leaving the remaining documents for training set (4/5) and validation set (1/5) for model selection

3 Metrics

Precision: fraction of positive predictions, actually labelled as positive

Recall: fraction of positive examples that are correctly detected

F1: harmonic mean between precision and recall

Baselines for comparison: random classifier

The experiments the team led followed the Leave-One-Out procedure.

Experimental Results

Performance: Training set size = 50 Tos

Method	P	R	F_1
SVM—single model	0.729	**0.830**	0.769
SVM—combined model	0.798	0.782	0.781
Tree kernels	0.777	0.718	0.739
Convolutional neural networks	0.729	0.739	0.722
Long short-term memory networks	0.696	0.723	0.698
SVM-HMM—single model	0.759	0.778	0.758
SVM-HMM—combined model	**0.859**	0.687	0.757
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

The best performing system is an ensamble

Preliminary results, obtained at the very beginning of the project with the first version of the dataset (50 ToS)

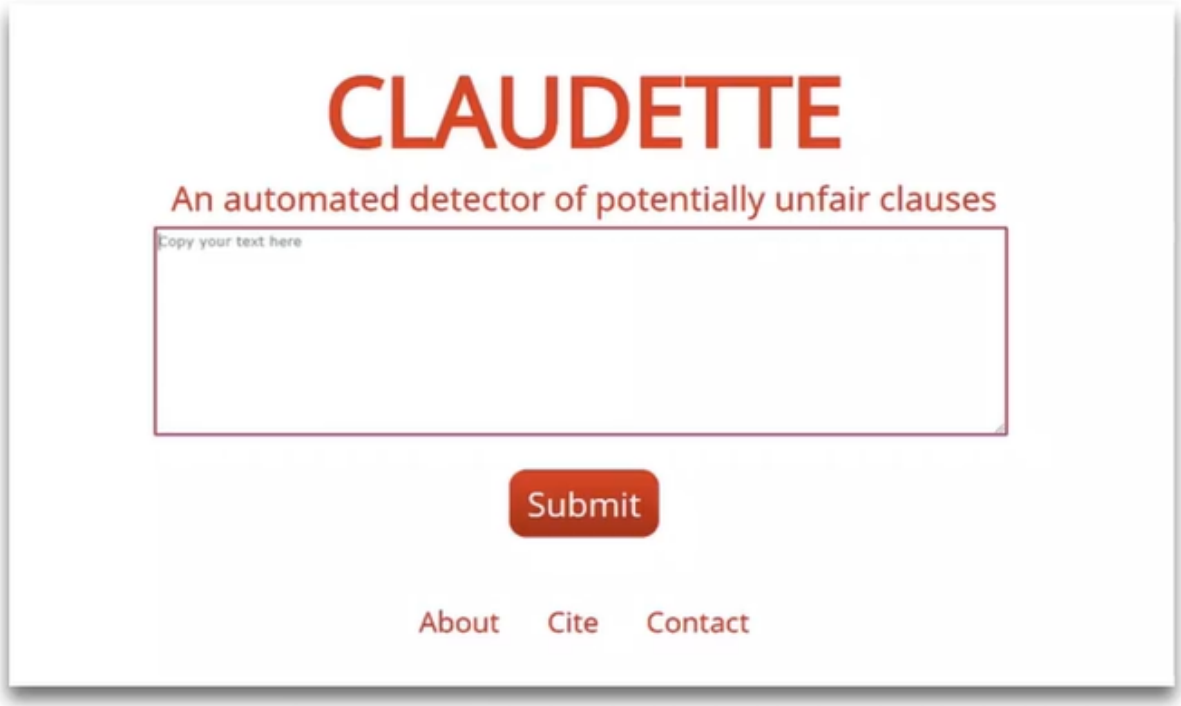
Experimental Results

Claudette correctly detected 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.

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

Some results are due to the heavy unbalance of the training set.

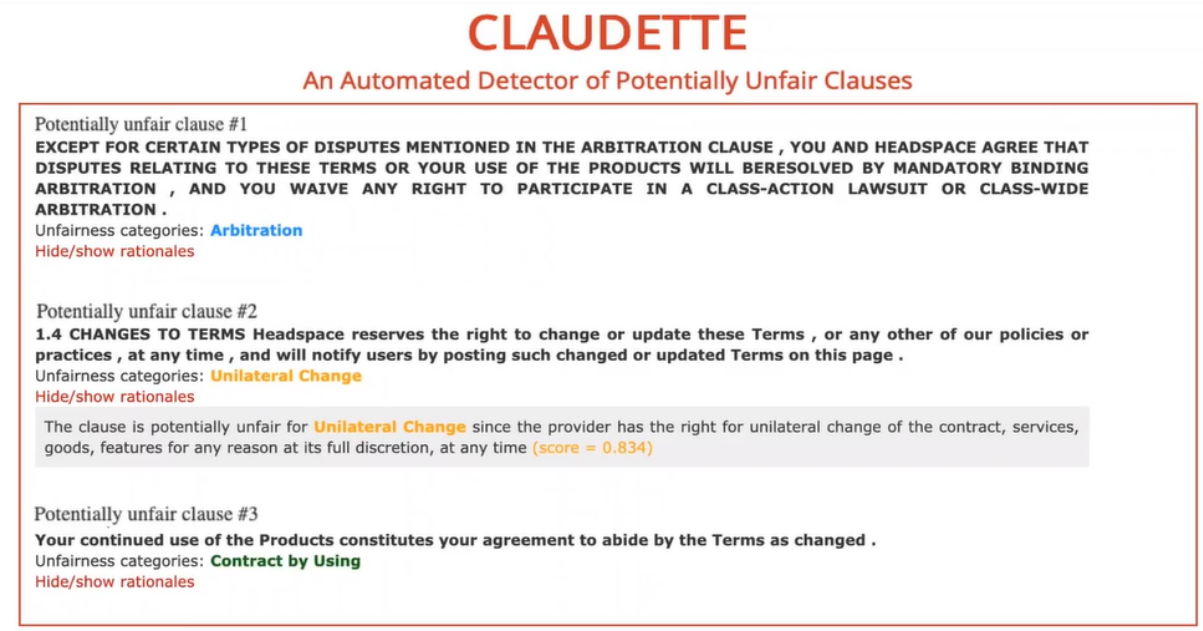
An online server



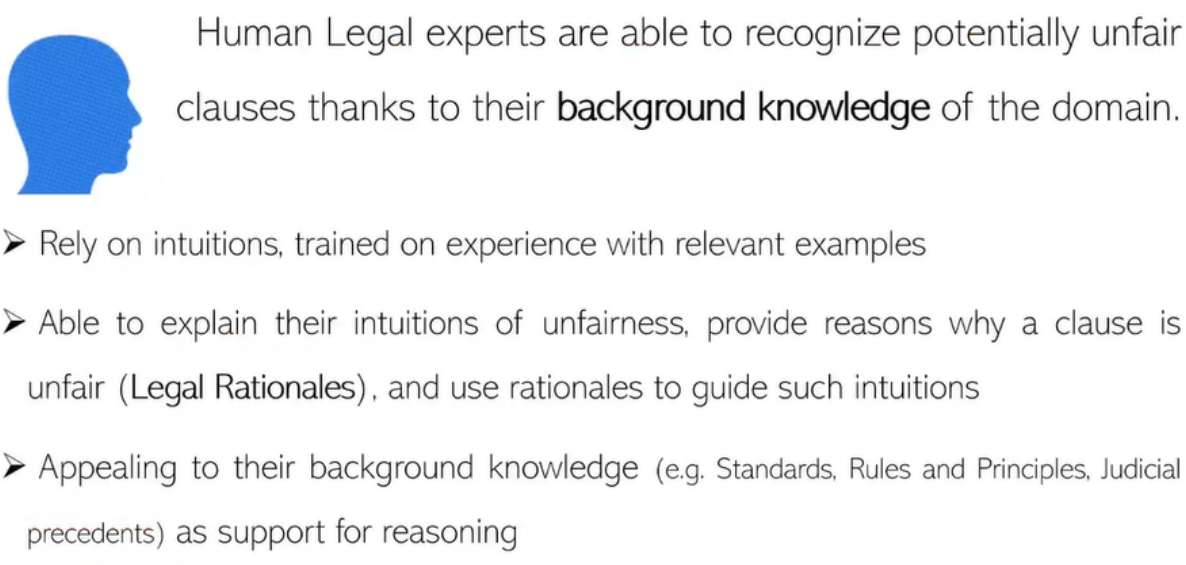
claudette.eui.eu

The idea of having an online server is that the user can simply copy paste any ToS and after submiting it, the system will assess its fairness.

Here we can see an example of *potentially unfair* clauses identified by CLAUDETTE and in the slide below it is also reported the **rationale** that the system used to come to that conclusion.



It also shows the score of each rationale: this **confidence** score is relevant for the system's **explainability**, how reliable the statement is and for the model's inspection.

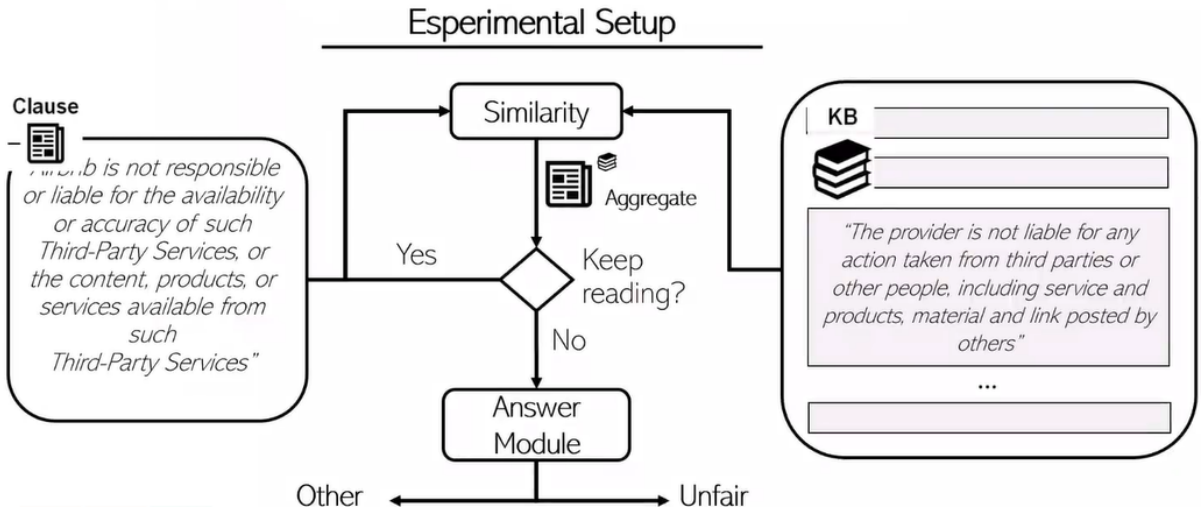


How could they achieve such results?

Memory-Augmented Neural Networks

- Process input and **store** the information in some **memory**
- Understand **pieces of knowledge** relevant to a given **query**
- Retrieve **concepts** from memory
- Combine **memory and query** to make a prediction

Exploiting Knowledge for Unfairness Identification



Graphic representation of how the unfairness detection problem has been formulated

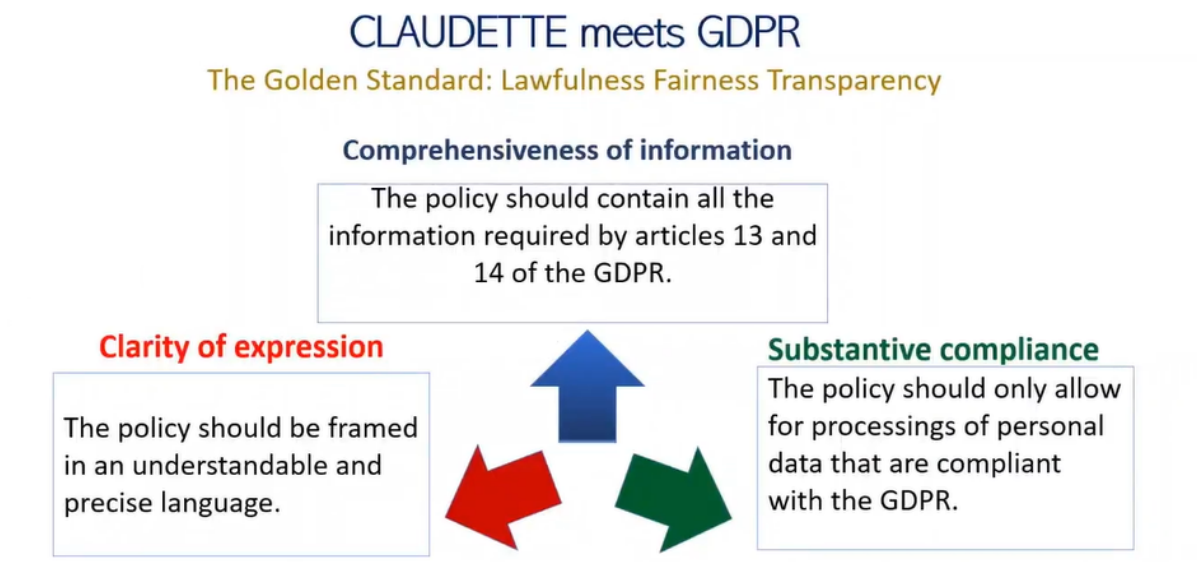
Firstly, they defined the **background knowledge** to be used by the system and as legal experts they defined a list of legal **rationales** (i.e. a list of possible explainations for unfairness) and associated to each an identifier. [Then each identifier can be linked to a clause considered *potentially or clearly unfair* contained in the dataset.]

Each clause in the dataset is given as an input to the system, which makes a query to the **knowledge base** and it compares the input clause to be assessed with those contained in the knowledge base. From this comparison, we extract the most relevant information through a **similarity score** of the two clauses. The idea is that if a clause is unfair then it might refer to a certain unfairness explanation for that

specific category in the memory. Then, the most relevant information (i.e. with the highest similarity score) is extracted from the memory and **aggregated** with the input, obtaining an ***enhanced input,*** which is used to repeat the same process multiple times, until all the relevant content has been extracted. This is done because the system might not be able to identify at first glance all the relevant information.

At the end of the aggregation phase, the system uses the information gathered to assess the clause fairness. So, there is an overlap between the rationales defined and the input clause that helps the system to classify the input.

For example, if we have a *limitation of liability* clause as input, it will have a low similarity score with a rationale about the *jurisdiction clauses*, which would be excluded from the aggregation.



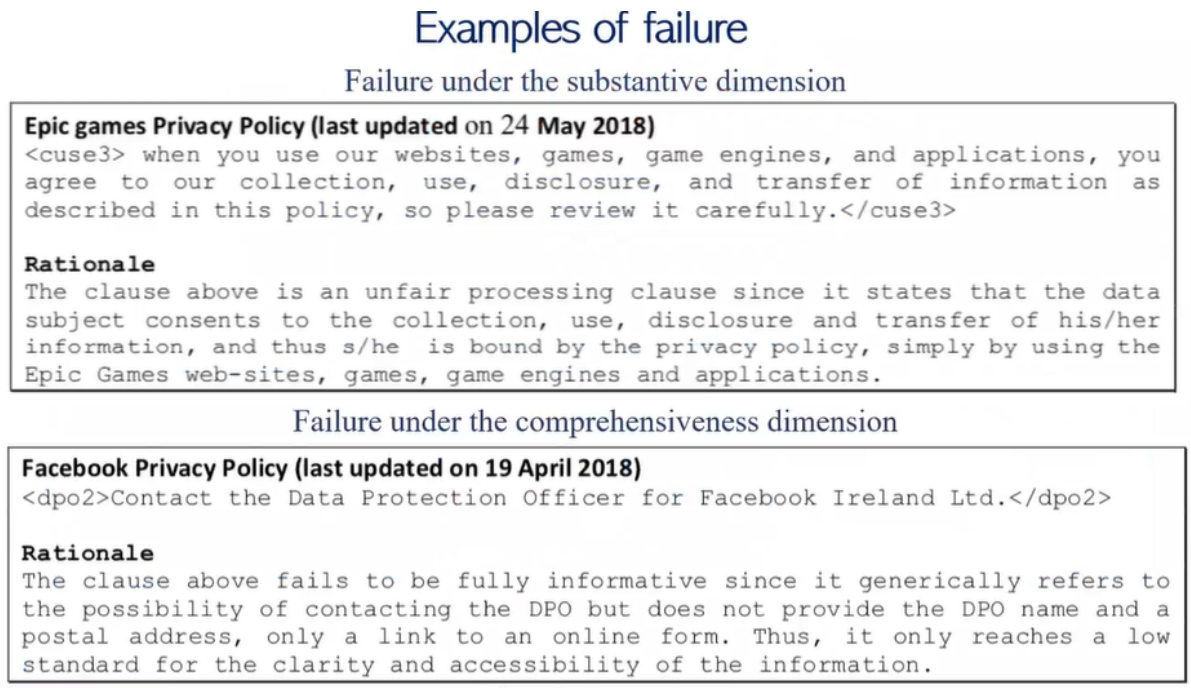
Different Levels of Achievement: Optimal and Suboptimal (questionable or insufficient)

This approach allowed to increase the system's performances. Given the good results on the ToS, the research group decided to extend the system focus to Privacy Policies and their compliance with the GDPR. The Golden Standard they defined contains how perfectly compliant PP should be in theory and they also defined 3 different dimensions.

- *Comprehensiveness of information:* all the information necessary to the user (e.g, the name of the controller, the purpose for which the data are collected and processed and so on).
- *Substantive compliance:* there should not be present unlawful practices.

- *Clarity of expression:* the PP should be framed in an understandable and precise language as it should not be possible to find different interpretations of the same clause.

For each of this dimensions they defined a list of category to look for and different possible levels of achievement.



The first clause is *clearly unfair* (differently from what it was considered in the ToS in the Consent by using category), because according to an article in the GDPR consent cannot be inferred by the use of a service but it should be given through **positive actions**.

The second clause fails to be comprehensive because it reports just one link as a contact, which is too limited since if we do not receive a reply we have no other means to reach the *Data Protection Officer (DPO)* who basically won't be accountable.

Comprehensiveness of information

23 categories (GDPR art 13 and 14)

Type of required information	Symbol
Identity of the controller (controller's representative)	<id>
Contact details of the controller (controller's representative)	<contact>
Contact details of the data protection officer	<dpo>
Purposes of the processing	<purp>
Legal Basis for the processing	<basis>
Categories of personal data concerned	<cat>
Recipients or categories of recipients of the personal data	<recep>
Period for which the personal data will be stored, or the criteria used to determine that period	<ret>
Right to lodge a complaint with a supervisory authority	<complain>
...	<...>

For each category has been defined an xml tag and a degree of optimality.

The Categories of personal data concerned

Clauses where clauses where the categories of personal data are comprehensively specified and not vague: fully informative

In other cases (e.g. when a clause only provides examples): insufficiently informative

Google Privacy Policy (last updated on 25 May 2018)

```
<cat1>We collect information about your location when
you use our services, which helps us offer features like
driving directions for your weekend getaway or showtimes
for movies playing near you.</cat1>
```

This clause is considered to be **fully informative** and is phrased in a good way because they specify **which** data they are collecting and the **purposes** of the collection (this part complies to the *substantive* requirements).

Often providers give out only a list of examples (usually not **exaustive)** of the data they are collecting or they define it by describing the interactions with the service used to collect data (which can mean basically anything). The result is an **insufficiently informative** clause.

Substantive compliance

10 categories (GDPR art 5, 6, 9 and others)

Type of clause	Symbol
Processing of special categories of personal data (e.g. health, sex life, political opinions, religious beliefs, etc.)	<sens>
Consent by using	<cuse>
Take or leave it approach	<tol>
Third party data transfers	<tp>
Policy change	<pch>
Transfer of data to third countries	<cross>
Processing of children's data	<child>
Licensing data	<lic>
Advertising	<ad>
Any other type of consent	<c>

Take it or leave it approach: you can either accept the use of your personal data or stop using the service. Here the consent cannot be considered as **freely given** (GDPR).

Policy change

When notice is given and new consent is required: fair processing clause

When notice is given but a new consent (or confirmation of reading) is not required: problematic processing clause

Twitter Privacy Policy (effective on 25 May 2018)

We may revise this Privacy Policy from time to time. The most current version of the policy will govern our processing of your personal data and will always be at https://twitter.com/privacy. If we make a change to this policy that, in our sole discretion, is material, we will notify you via an @Twitter update or email to the email address associated with your account.

Differently from what we have seen in the ToS, which are contracts, the PPs under the GDPR have the **duty to inform** the user and whenever they change, the user should give her/his consent **again**. When this requirement is not met, and consent is the **legal base** for the processing of a specific kind of data,the clause has been considered as *problematic.*

The clause above is problematic because the DPO is the one who decides in his *sole discretion* whether or not the changes to the policy affect the user and if he/she should be notified (problematic w.r.t. the *substantive compliance dimension).*

Policy change

When no notice is given and new consent is not required: unfair processing clause

Booking Privacy Policy (last updated on 9 May 2018)

We might amend the Privacy Statement from time to time. If you care about your privacy, visit this page regularly and you'll know exactly where you stand.

Cattiva questa.

Clarity of expression

(GDPR art. 5(1)(a), 12 (1) and others)

Is the privacy policy framed in an understandable and precise language?

4 main indicators of vagueness

Indicator	Language qualifiers
1. Conditional Terms The performance of a stated action or activity is dependent on a variable trigger	Depending, as necessary, as appropriate, as needed, otherwise reasonably, sometimes, from time to time, etc.
Example	**Rationale**
<vag> We also may share your information if we believe, in our sole discretion, that such disclosure is **necessary:**…."</vag>	The practice described as "necessary" suggests that the sharing will only occur in exceptional cases, however the clause fails to specify under what exceptional conditions the provider will disclose the information.

Extremely Vague

Clarity of expression

Indicator	Language qualifiers
2. Generalization: i.e. terms that vaguely abstract information practices using contexts that are unclear. Action(s)/Information Types are vaguely abstracted with unclear conditions.	generally, mostly, widely, general, commonly, usually, normally, typically, largely, often, primarily, among other things, etc.
Example	**Rationale**
<vag> We **typically** or **generally** collect information …</vag> <vag> When you use an Application on a Device, we will collect and use information about you in **generally** similar ways and for similar purposes as when you use the TripAdvisor website.</vag>	The use of the generalization term "generally" obscures for the data subject the service provider activities, since it provides a large flexibility to the service provider.

If the data controller decides to process the data for a different purpose and he believes that is more or less *generally* similar to other (allowed?) purposes, then the user has not the right to be informed about those practices.

Clarity of expression

(GDPR art. 5(1)(a), 12 (1) and others)

Indicator	Language qualifiers
3. Modality: it includes modal verbs, adverbs and non-specific adjectives, which create uncertainty with respect to actual action; it includes whether an action is possible. Modality does not include whether an action and/or activity is permitted. Modality mainly refers to ambiguous possibility of action or event.	may, might, could, would, possible, possibly, etc.
Example	**Rationale**
<vag>We **may** use your personal data to develop new services </vag>	it is unclear whether or not the controller will use the data subject information.to develop new services and in what cases and under

Clarity of expression

Indicator	Language qualifiers
4. Non specific Numeric quantifiers: which create ambiguity as to the actual measure	certain, numerous, some, most, many, various, including (but not limited to), variety
Example	**Rationale**
<vag>When you create an Apple ID, apply for commercial credit, purchase a product, download a software update, register for a class at an Apple Retail Store, connect to our services, contact us or participate in an online survey, we may collect a **variety** of information, **including** your name, mailing address, phone number, email address, contact preferences, device identifiers, IP address, location information and credit card information.</vag>	it creates ambiguity with regard to the actual measure of information the data controller collect

The lists of qualifiers above are **indicators** to focus the attention of the taggers, and possibly the system, to examine whether or not the clause is vague.

Clarity of expression

A combination of different forms of vagueness

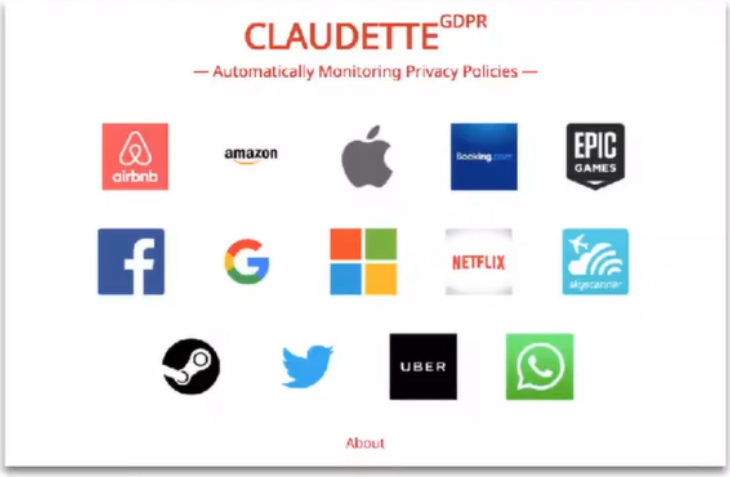
<vag>We **generally may** share personal information we collect with **certain** service providers, **some** of whom **may** use information for their own purposes **as necessary.**</vag>

In combination, these six forms of vagueness allow any organization sharing personal information under this statement to share it with anyone for any purpose, as long as the recipient is a service provider. The conditions under which information is shared, and the number or proportion of service providers that engage in this practice remain unclear.

This clause conveys no useful information, is basically contentless.

CLAUDETTE FOR GDPR

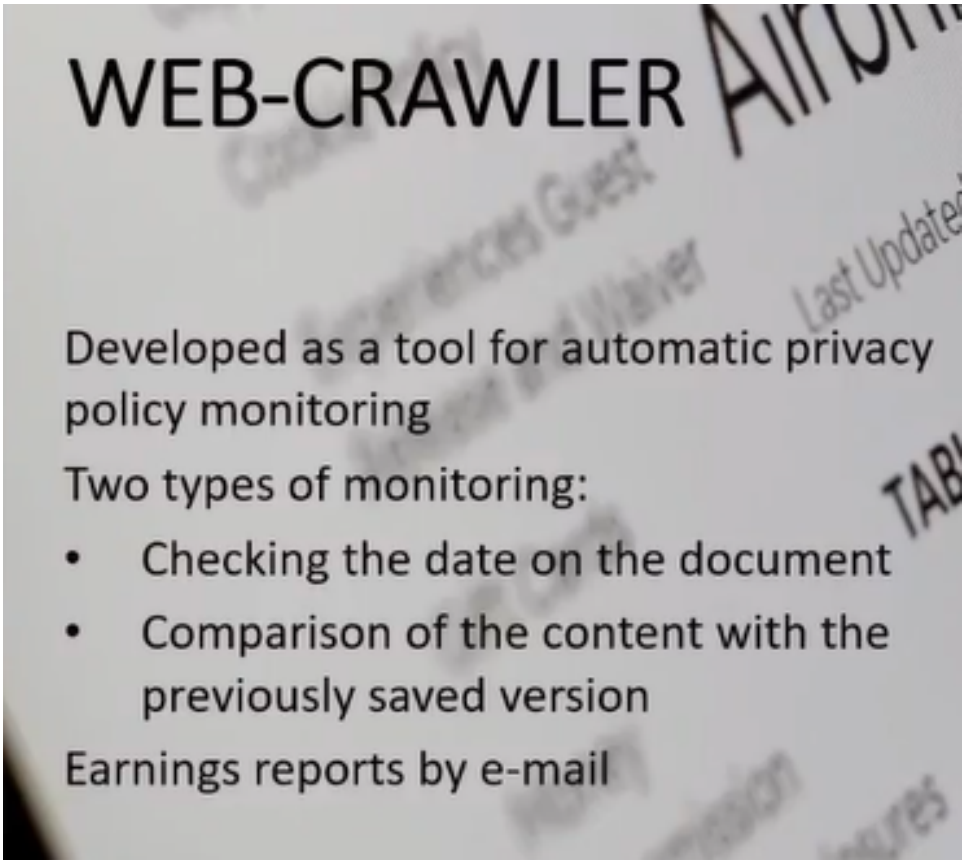
http://claudette-gdpr.eu



WHERE DID WE START?

- 14 documents (100 now)
- 3,658 sentences
- 80,398 words
- 11.0% sentences contain unclear language
- 33.9% sentences contain potentially unlawful clauses

The assessment of privacy policy is more complex and the result have not been as good as the ones obtained with ToS. So they are trying different algorithms that can take into account the context, because the PP's frame is articulated throughout different sentences, while ToS's sentences can be considered on their own.



What we are working on …

- Experimenting new method for privacy policies
- Multilingualism (The Claudette german version)
- Empowerment through transparency
  - Linguistic transparency
  - Provide explanations opening black box AI Systems