AI Researchers Estimate 97% Of EU Websites Fail GDPR Privacy Requirements- Especially User Profiling

Originally published at https://www.unite.ai/ai-researchers-estimate-97-of-eu-websites-fail-gdpr-privacy-requirements-especially-user-profiling/

Researchers in the US have used machine learning techniques to study the GDPR privacy policies of over a thousand representative websites based in the EU. They found that 97% of the sites studied failed to comply with at least one requirement of the European Union's 2018 regulatory framework, and that they complied least of all with regulatory requirements around the practice of 'user profiling'.

The paper states:

> *'[Since] the privacy policy is the essential communication channel for users to understand and control their privacy, many companies updated their privacy policies after GDPR was enforced. However, most privacy policies are verbose, full of jargon, and vaguely describe companies' data practices and users' rights. Therefore, it is unclear if they comply with GDPR.'*

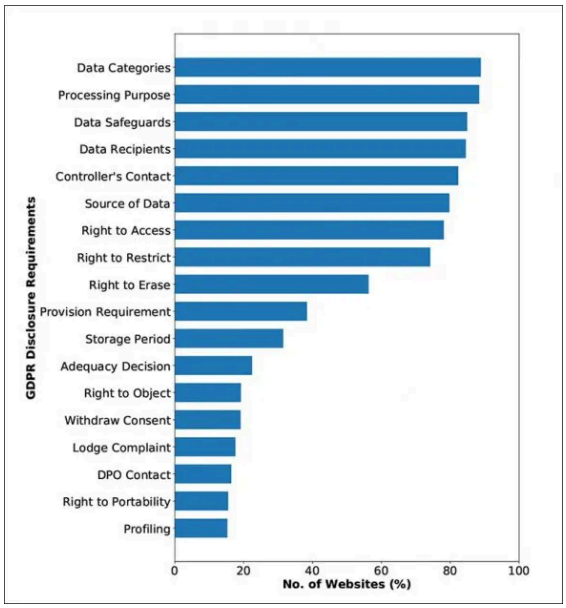
It continues:

> *'Our results show that even after GDPR went into effect, 97% of websites still fail to comply with at least one requirement of GDPR.'*

The study is titled *Automated Detection of GDPR Disclosure Requirements in Privacy Policies using Deep Active Learning*, and comes from three researchers at the University of Virginia at Charlottesville.

Privacy Last

The area of least compliance, according to the study, concerned GDPR's stipulations about user profiling, with the authors stating that only 15.3% of the sites studied were in full compliance with this particular rule.



A graph of compliance among websites' GDPR policies studied for the research. Source: https://arxiv.org/pdf/2111.04224.pdf

User profiling (where a person's interaction with websites is recorded and often used to 'target' them in other online contexts, such as advertising) has become one of the hottest controversies in tech since the Cambridge Analytica scandal.

On Tuesday, a key committee of the European Parliament passed the first stage of the new Digital Markets Act (DMA) legislation, which would ban the behavioral targeting of minors, imposing fines of up to 20% of *global* annual sales for infringing companies.

Though the Act has been received by the media as a direct response to the growing influence of tech giants such as Facebook and Google, the sheer scale of non-compliance represented by the new research suggests that the vast majority of EU companies (including EU-resident offices for American companies trading in Europe) are legally exposed to GDPR fines.

Additionally, Italy has this week imposed the maximum allowable fine of 10 million euros ($11.2 million USD) against Apple and Google for exploiting user profiling, among other infractions.

Data

The sites examined in the new research were sampled from the top 10,000 websites listed in Quantcast, the English-language privacy policies of which were extracted through Yandex searches on UK-based VPNs (in order to ensure that the policies were not geo-blocked).

EU websites have been obliged to provide prescribed privacy policies, covering 18 central requirements (see graph above) since the General Data Protection Regulation (GDPR) act came into full effect in May 2018.

The researchers limited their extraction of privacy policies to a period from August 2018 onward, to allow reasonable time for domains to have published the required policies (a requisite that they had advance knowledge of for at least a year of the two-year development phase of GDPR since 2016).

The filtering process produced a privacy corpus of 9,761 policies, from which 1,080 policies were randomly selected by the researchers.

Pre-Processing

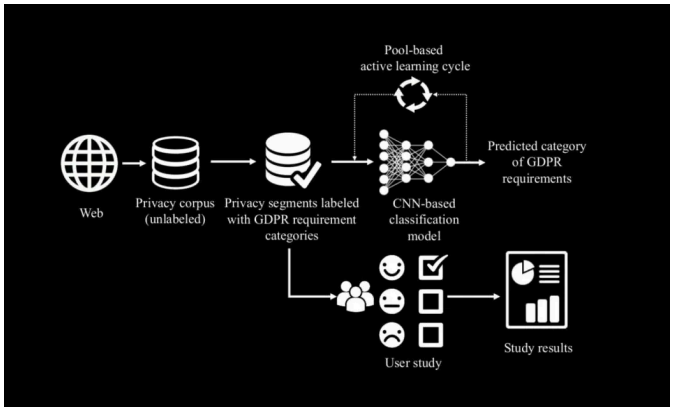
The team employed two legal experts to train four human annotators to label each of the 18 possible privacy policies mandated by GDPR.

Some of the legalese in the policies covered more than one of the 18 requirements, making it necessary to use a Convolutional Neural Network (CNN) to detect language features associated with each policy.

An initial attempt to train a model to identify compliance based on language achieved 80.5% success. To improve these results, the researchers applied Active Learning to bolster the model's performance using less labeled data. By these means it was possible to train the classifier CNN up to an accuracy of 89.2%, with an F1 score of 0.88 (where '1' is complete success).

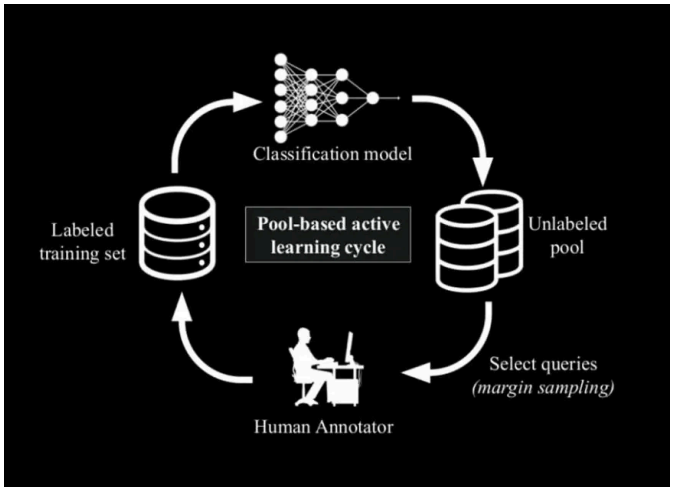
To ensure the word embeddings were specific to privacy policy, the researchers trained an unsupervised word embedding model using Facebook's FastText Python library.

As per standard practice, the final data was split 80/20 between trained data and test data (i.e. randomly selected data against which the accuracy of the algorithm will be judged). A human-in-the-loop measurement study was added to the architecture in order to evaluate the quality of results.



The architecture for the classifier system.

In the course of the workflow, 11,271 human-annotated privacy policy segments were produced, each of which was reviewed by four human annotators that had been trained by the two legal experts involved in the study. Where disagreement occurred, a 75% agreement ratio was needed in order not to reject the data from inclusion.



Humans-in-the-loop – it was not possible to entirely automate the labeling of the policy data, though Active Learning enabled a pool-based workflow that made the project feasible.

Besides the results already mentioned, the users found that *portability* – the right under GDPR to translocate or export data held by a company – was almost as poorly served as profiling.

The researchers conclude:

> *'[Requirements] such as users' Right to Portability and providing the contact information of Data Protection Officer (DPO contact) are covered by 15.5% and 16.4% websites, respectively. Other primary requirements, such as users' right to Lodge Complaint, Withdraw Consent, Right to Object, and Adequacy Decision, are covered by17-20% websites.'*

…and continue:

> *'It appears that only 3% of websites fully comply with 18 requirements. These findings indicate that many websites still do not follow the requirements of GDPR.'*

7pm 26/11/2021 – Clarified first graph caption. – MA

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More