applied sciences

MDPI

Article

Automatic Assessment of Privacy Policies under the GDPR

David Sánchez *, Alexandre Viejo and Montserrat Batet

Center for Cybersecurity Research of Catalonia (Cybercat), Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, UNESCO Chair in Data Privacy, Av. Països Catalans, 26, E-43007 Tarragona, Spain; alexandre.viejo@urv.cat (A.V.); montserrat.batet@urv.cat (M.B.)

* Correspondence: david.sanchez@urv.cat

Abstract: To comply with the EU General Data Protection Regulation (GDPR), companies managing personal data have been forced to review their privacy policies. However, privacy policies will not solve any problems as long as users do not read or are not able to understand them. In order to assist users in both issues, we present a system that automatically assesses privacy policies. Our proposal quantifies the degree of policy compliance with respect to the data protection goals stated by the GPDR and presents clear and intuitive privacy scores to the user. In this way, users will become immediately aware of the risks associated with the services and their severity; this will empower them to take informed decisions when accepting (or not) the terms of a service. We leverage manual annotations and machine learning to train a model that automatically tags privacy policies according to their compliance (or not) with the data protection goals of the GDPR. In contrast with related works, we define clear annotation criteria consistent with the GDPR, and this enables us not only to provide aggregated scores, but also fine-grained ratings that help to understand the reasons of the assessment. The latter is aligned with the concept of explainable artificial intelligence. We have applied our method to the policies of 10 well-known internet services. Our scores are sound and consistent with the results reported in related works.

Keywords: privacy policies; GDPR; privacy goals; privacy assessment; machine learning

Citation: Sánchez, D.; Viejo, A.; Batet, M. Automatic Assessment of Privacy Policies under the GDPR. *Appl. Sci.* **2021**, *11*, 1762. https://doi.org/10.3390/app11041762

Academic Editor: José A. Ruipérez-Valiente

Received: 8 January 2021

Accepted: 13 February 2021

Published: 17 February 2021

1. Introduction

FaceApp [1]—the Russian face-aging app—went viral in July of 2019 by means of the FaceApp and the Old-Face Challenges, which were cheerfully followed by celebrities, influencers, and common people. As a result, over 150 million people downloaded the application from Google Play and freely provided their names and photos to the company behind the app [2]. The fact that a new app emerges, goes viral, and millions of people download it is not especially remarkable. Nevertheless, in the FaceApp case, there was a paragraph buried among the app's terms of service that raised the scandal (see the terms of service published in 08/03/2017 [1]):

> *"You grant FaceApp a perpetual, irrevocable, nonexclusive, royalty-free, worldwide, fully-paid, transferable sub-licensable license to use, reproduce, modify, adapt, publish, translate, create derivative works from, distribute, publicly perform and display your User Content and any name, username or likeness provided in connection with your User Content in all media formats and channels now known or later developed, without compensation to you."*

The millions of users that installed the app accepted these terms—and several other statements as threatening as this one. These users might not have read such terms or might not have been aware of what they were accepting. Whatever the case may be, the company became full right perpetual owner of millions of photos tagged with the corresponding true identities [1].

In the last years, awareness of both the privacy threats and the enormous economic value of personal data has risen significantly. Governments around the globe have been

striving to give people tools to protect their rights against the entities that try to collect and monetize their personal data. The recent EU General Data Protection Regulation (GDPR) is probably the most significant materialization of this intention. To comply with the GDPR, companies managing personal data have been forced to review and clarify their privacy policies in order to make the users aware of how their data are collected and used. However, scandals such as the one starring by FaceApp show that the existence of privacy policies (including abusive ones) will not solve any problem as long as the users do not read them [3] or are not able to understand what they are accepting. As shown in [4], the reasons for the latter are: (i) most privacy policies take between 20 to 25 minutes to read; and (ii) most policies are too technical and obscure to be understandable by non-technical users.

In order to address these issues, in this paper we present a system that automatically assesses privacy policies under the umbrella of the GDPR. In particular, our proposal quantifies the degree of policy compliance with respect to the data protection goals stated by the GPDR and presents clear and intuitive privacy scores to the user. As a result, any user (even those who are not familiar with the GDPR), will become immediately aware of the risks associated with the offered services and their severity; this will empower them to take informed decisions when accepting (or not) the terms of a service. Our proposal leverages manual annotations by privacy experts and machine learning to train a model that automatically tags privacy policies according to whether or not they comply with the data protection goals of the GDPR. From this, we derive global and per-goal privacy scores that explain the results of the policy assessment.

The remainder of the paper is organized as follows. Section 2 discusses related works and highlights their limitations. Section 3 gives an overview of the GDPR and characterizes its data protection goals. Section 4 details how we designed and trained our policy assessment model. Section 5 explains the policy scoring we propose. Section 6 reports empirical experiments on the policies of 10 well-known internet services and discusses the results. The final section gathers some concluding remarks and depicts lines of future research.

2. Related Works

We can find few proposals in the literature that tackle this issue. All of them follow an approach similar to ours: they employ natural language processing, manual annotation, and machine learning to automatically tag privacy policies. However, they also use ad-hoc annotation categories that do not match or just partially cover the data protection goals stated in the GDPR.

Specifically, in [5], the authors annotate policy statements according to items such as "Collection of Personal Info" or "Personal Info is stored for Unlimited Time". In [6], the authors tag statements according to categories such as "First Party Collection/Use", "Third Party Sharing/Collection", "Policy Change" or "Do Not Track". The system presented in [7] answers questions such as "Does this website's privacy policy disclose whether data it collects about you is used in ways other than you would reasonably expect given the site's service?" or "Does this website's privacy policy disclose whether the site or service tracks a user's actual geolocation?". Finally, the system proposed in [8] also provides answers to 10 questions regarding data management such as "How does this site handle your email address?" or "Does this site track or share your location?".

In [9], the authors annotate policies with regard to 10 privacy aspects derived from pre-GDPR works. Even though some of these are similar or partially overlap with the privacy goals we consider in our work, the annotation criteria are not detailed and a filtering step is introduced to consider only the sentences containing a predefined set of keywords. Even though the latter simplifies the classification, it may also omit a significant number of sentences that could contribute towards a more accurate characterization. In [10], the authors present a dataset of privacy policy snippets annotated according to five handpicked privacy requirements. However, since the work focuses on measuring the performance of different classifiers, no privacy scores were derived from such annotations.

Guard (https://useguard.com/ (accessed on 16 February 2021)), being a commercial product, is opaque regarding how statements are annotated: it simply grades privacy policies with a mark (being A+ the most privacy friendly, and F the worst) calculated from the number of generic privacy threats found in the policy and the number of scandals in which the company behind the policy has been involved.

The results offered by these approaches either offer a partial view of the privacy goals stated in the GDPR or merely provide aggregated scores; even though they may allow ranking and comparison of policies, they do not explain the reasons and threats underlying to such scores.

3. The GDPR Privacy Requirements and the Data Protection Goals

The legal framework defined in the GDPR lays down fundamental rules for the protection of personal data. In particular, Article 5 lists the essential requirements on the security of the processing of personal information that data controllers and service providers should consider in their data processing procedures:

- Transparency for the data subject affected by the processing of personal data.
- Purpose limitation for the processing of personal data.
- Data minimization in the processing of personal data.
- Accuracy of personal data.
- Storage limitation for personal data.
- Integrity of personal data.
- Confidentiality of personal data.
- Accountability and verifiability.

Moreover, the GDPR recognizes several rights of the data subjects that the controller must grant through technical and organizational measures:

- Support in exercising the rights of data subjects (Article 12).
- Identification and authentication of the person requesting information (Article 12).
- Right to rectification (Article 16).
- Right to erasure (Article 17).
- Restriction of data processing (Article 18).
- Data portability (Article 20).
- Possibility to intervene in processes of automated decisions (Article 22).
- Freedom from error and discrimination in profiling (Article 22).

Finally, the GDPR also states several technical requirements to promote data protection:

- Data protection by default (Article 25).
- Availability of systems, services and data (Article 32).
- Resilience of the systems and services (Article 32).
- Restorability of data and data access (Article 32).
- Evaluability (Article 32).
- Rectification and mitigation of data protection violations (Articles 33 and 34)
- Adequate monitoring of the processing (Articles 32, 33 and 34).

To facilitate their materialization into technical and organizational measures, the legal requirements of the GDPR have been structured into the following data protection goals [11]:

- *Data minimization* requires no collection, processing, and use of more personal data than necessary for the achievement of the purpose of the processing. This principle assumes that optimal data protection is achieved when no or as little personal data as possible are collected and processed. This goal encompasses the following GPDR requirements: *data minimization*, *storage limitation*, *data protection by default*, and *evaluability*.
- *Availability* requires the collected personal data to be accessible to authorized parties. This goal encompasses the following GPDR requirements: *availability*, *resilience*, *restorability*, *evaluability*, and *rectification and mitigation of data protection violations*.

- *Integrity* requires that the data to be processed remain intact, complete, and up-to-date. This goal encompasses the following GPDR requirements: *accuracy*, *integrity*, *freedom from error and discrimination in profiling*, *resilience*, *rectification and mitigation of data protection violations*, *evaluability*, and *adequate monitoring of the processing*.
- *Confidentiality* requires that no person is allowed to access personal data without authorization. Non-authorized persons include third parties external to the controller, employees who do not need access to personal data for the provision of the service, and, in general, anyone unrelated to the data processing purpose for which consent of the data subjects was obtained. This goal encompasses the following GPDR requirements: *confidentiality*, *resilience*, *rectification and mitigation of data protection violations*, and *evaluability*.
- *Unlinkability* requires data to be processed and analyzed only for the purpose for which they were collected, and prohibits combining data with other sources to derive more detailed records. This goal encompasses the following GPDR requirements: *purpose limitation* and *evaluability*.
- *Transparency* requires the data subject to be able to understand which data are collected and processed for a particular purpose, which systems and processes are used for this purpose and who is legally responsible for the data and systems in the various data processing steps. This goal encompasses the following GPDR requirements: *transparency*, *accountability and verifiability*, *evaluability*, and *adequate monitoring of the processing*.
- *Intervenability* requires the data controller to effectively grant the data subjects their rights to notification, information, rectification, blocking, and erasure at any time. This goal encompasses the following GPDR requirements: *support in exercising the rights of data subjects*, *identification and authentication*, *right to rectification*, *right to erasure*, *restriction of data processing*, *data portability*, *possibility to intervene in processes of automated decisions*, *data protection by default*, *evaluability*, and *rectification and mitigation of data protection violations*.

4. Learning the GDPR Data Protection Goals

Our approach assesses privacy policies under the framework defined by the data protection goals enounced above. Roughly, the approach considers a privacy policy to be "good" if it provides explicit evidence that the measures required to encompass such goals are undertaken by the service provider. On the contrary, it considers a policy to be "bad" if the way in which personal data are collected, processed, and managed by the service provider contradicts the principles and requirements of the data protection goals. Moreover, our approach considers ambiguous statements, lack of detail, or omissions in any relevant aspect of data management to be "bad" as well. It is worth mentioning that this latter type of statement is more common than those explicitly contradicting data protection goals. Since the assessment is done on a goal basis, our system is capable of both qualifying the policy as a whole and characterizing it according to its degree of compliance with respect to each goal.

We approach the assessment of policies as a multi-label text classification task. This is, for a given policy, each statement (sentence) in the policy is classified (tagged) according to whether or not it complies with each data protection goal. For this purpose, we defined the following 14 tags: *data minimization compliant*, *data minimization non-compliant*, *availability compliant*, *availability non-compliant*, *integrity compliant*, *integrity non-compliant*, *confidentiality compliant*, *confidentiality non-compliant*, *unlinkability compliant*, *unlinkability non-compliant*, *transparency compliant*, *transparency non-compliant*, *intervenability compliant*, and *intervenability non-compliant*.

Then, we trained a support vector machine (SVM) to automatize the classification of statements. An SVM is a fast, supervised classification algorithm that performs well with a limited amount of training data. Indeed, with a limited number of training samples, SVMs tend to offer better performance than "modern" machine learning algorithms, such as convolutional neural networks [12]. This makes them suitable for our setting, because

the text available in privacy policies contains just few hundred sentences rather than the millions of records typically considered in deep learning tasks. For the SVM to work, we treat each sentence in the policy as a bag-of-words and feed it into the SVM as a vector of features, each one representing the relative frequency of appearance of each word in the sentence. Vectors have as many features as words in the vocabulary defined by the collection of texts used as training data.

We also reduce the size of the vocabulary by removing stop words and by using a stemming algorithm to consider morphological variations of the same word as equivalent. Additionally, instead of treating with isolated words, we employed a pipeline of linguistic analyses to tokenize, part-of-speech tag, and chunk text. In this way, vocabulary entries for the SVM consist of phrases (or n-grams) rather than of words. Using n-grams also helps to improve the accuracy of the classification because phrases are less ambiguous than isolated words (e.g., "family name" vs "family" + "name"). Even though this pre-processing helps to decrease the vocabulary size, it is common to end up with vectors having thousands of features, one for each distinct phrase in the training data set. For such high-dimensional scenario, SVMs perform better when using standard linear kernels to separate classes because non-linear ones tend to overfit [13].

We trained the SVM with manually tagged sentences extracted from the privacy policies of Google (https://policies.google.com/privacy?hl=en-US (accessed on 16 February 2021)), Mozilla (https://www.mozilla.org/en-US/privacy/ (accessed on 16 February 2021)), Netflix (https://help.netflix.com/legal/privacy (accessed on 16 February 2021)) and Twitter (https://twitter.com/en/privacy (accessed on 16 February 2021)). We selected these companies due to them being among the most relevant data controllers worldwide and because they have significantly different perspectives on data protection. In all cases, we considered the English versions of their policies.

Each of the authors of this work independently annotated each statement (sentence) in the training data set. To guide the work, we defined the following criteria to annotate sentences, which aim at capturing the essence of the data protection goals (and the underlying privacy requirements) represented by each annotation tag:

- *Data minimization*. A statement is tagged compliant with data minimization if the need of collecting and storing a personal attribute is justified strictly in the context of the offered service. Policies should detail the exact attributes collected, their need for the services offered to the user, and justify that data will only be made accessible to the entities involved in the service delivery. Positive statements are also those that describe data removal measures once the data are no longer needed or the user stops using the service. On the contrary, a non-compliant statement is that which does not properly justify the need of data collection, does not detail the collected attributes, or seems to store users' data indefinitely.
- *Availability*. To be compliant, the policy should explicitly state that users have the means to access the data collected on them. This may also include measures to ensure that this access is possible under any circumstances, including in the event of security attacks and system failures. The user should also have the means to contact the company at any time.
- *Integrity*. Compliant statements are those that detail security measures implemented by the service provider to ensure that the data will not be modified due to external factors, such as security attacks or system failures.
- *Confidentiality*. Any statement describing the collection and storage of clear personal attributes—regardless of their secondary use—is considered negative for confidentiality because it may entail privacy leakages. Statements that detail appropriate measures to control the access to the collected data and/or to ensure their anonymity are considered positive. Sound anonymity measures implemented at the server side that include storing only aggregated data of several users, replacing attribute identifiers by pseudonyms, or limiting the level of detail of the collected data are considered positive as well, because this will minimize the harm in case of data leakages.

- *Unlinkability*. Any statement describing the collection of identifying attributes—e.g., the name of the user, her IP, her device's hardware identifiers, etc.—or the use of tracking mechanisms (such as cookies or GPS tracking) are considered negative for unlinkability because they may be used to link and aggregate the users' data. Statements suggesting the possibility of combining data from different sources—either internal or external to the service provider—are also considered negative. Positive statements are those that explicitly refrain from collecting such attribute types or those that allow the user to access the services anonymously, that is, without registering and logging into the system.
- *Transparency*. Statements describing any aspect of the data collection, management, or use with enough detail and/or with proper justification are considered good for transparency. This also includes references to resources describing the company's data management activities and clear contact points to attend requests for clarifications or claims by the users. Statements that might be negative regarding other goals such as confidentiality or data minimization may be considered good for transparency if they are described with enough detail; for example, by providing specific lists of the personal attributes being collected. Negative statements for transparency are those that (i) do not detail the concrete attribute values involved in data collection or management—e.g., they talk about "data" in general; (ii) do not justify the need or uses of the data; or (iii) use ambiguous constructions, such as "we may" or "under some circumstances", among others.
- *Intervenability*. Positive statements from the intervenability point of view are those that give control to the users regarding granting or not their data to the service provider and rectifying and erasing such data at any time. Pro-active notifications regarding changes in the policy or about anything relevant on the users' data are considered positive as well. Negative statements are those that depict unilateral or automatic actions or algorithms implemented by the service provider that do not allow the intervention of the users. Finally, the lack of user control on any of the aspects detailed before is considered negative as well.

After annotating the statements independently, we reached an agreement on the divergent annotations. The resulting agreed tags constituted the training data set. To give some context to the annotation process, we computed our inter-annotator agreement as the average Fleiss' kappa between the 14 tags, which scored 0.68. Fleiss' kappa is a statistical measure for assessing the reliability of agreements between any number of annotators when assigning categorical ratings to a number of items. It is generally agreed that it provides a more robust measure than simple percent agreement calculation because it takes into account the possibility of the agreement occurring by chance. Fleiss' kappa values above 0 indicate positive agreement, values above 0.41 are considered moderate agreement, above 0.61 are considered substantial agreement and above 0.81 are considered almost perfect agreement. The inter-annotator agreement in our case is in the lower range of the substantial interval. Taking into account that we are data privacy experts who followed a common annotation criteria, we would like to stress two significant thoughts: (i) multi-label privacy-oriented annotation of policies is a difficult task; and (ii) privacy policies are complicated legal texts full of ambiguous statements that can be subjected to different interpretations. This is consistent with recent studies concluding that most privacy policies are unnecessarily vague and verbose [14], full of legal jargon and opaquely and ambiguously establish companies' justifications for collecting and selling data [4].

The training data set consisted of 488 tagged sentences, which constitute a random subset of the statements contained in the policies we considered. Note that some sentences may not have any tags if they do not refer to any aspect of personal data management. Sentences may have 7 tags at most, because our 14 tags are pairwise mutually exclusive. Most sentences had 1 or 2 tags since they usually cover a specific aspect of the data management. The total number of tags was 735, whose distribution is detailed in Table 1.

Table 1. Distribution of tags in the training data set

Tag	Number (Percentage)
data minimization compliant	14 (1.9%)
data minimization non-compliant	38 (5.17%)
availability compliant	29 (3.95%)
availability non-compliant	0 (0%)
integrity compliant	19 (2.59%)
integrity non-compliant	0 (0%)
confidentiality compliant	68 (9.25%)
confidentiality non-compliant	109 (14.83%)
unlinkability compliant	12 (1.63%)
unlinkability non-compliant	83 (11.29%)
transparency compliant	150 (20.41%)
transparency non-compliant	45 (6.12%)
intervenability compliant	107 (14.56%)
intervenability non-compliant	61 (8.3%)

We can see that, whereas some goals such as *confidentiality* and *transparency* were well covered, others were barely referenced, specifically *availability* and *integrity*. These two goals are more related to security than to privacy and, in fact, whereas privacy-related goals such as *confidentiality* and *transparency* may represent a conflict of interest between users and companies, ensuring security is usually in the company's own interest. This explains why they are not referenced very often in the privacy policies and, when done, they are on the positive side. Considering these arguments and, especially, the lack of negative samples for such goals, we decided to drop both the *availability* and *integrity* goals in the subsequent study.

5. Scoring Privacy Policies

Once trained, the SVM provides for each statement a set of tags and a confidence score (classification threshold) that measures the relative distance of a sample with respect to the hyperplane separating the classes learnt by the SVM. We consider a tag to be reliable if the confidence score is 60% or higher. Specifically, by fixing a value above 50% we focus on tags that exhibit a dominant class membership.

To score the degree of policy compliance p for a given goal i we measured the difference between the number of positive and negative tags obtained for that goal i. If the difference is positive (negative) for a given goal, we consider that the policy is compliant (non-compliant) with such goal in a degree proportional to the magnitude of the difference. Scores were normalized by the number of tags associated with the policy for all goals (G) and expressed as percentages. In this way, one can directly compare policies regardless the number of tags they have associated.

$$Score\_goal_i(p) = \frac{\#positive\_tags_i(p) - \#negative\_tags_i(p)}{\sum_{g \in G}\left(\#positive\_tags_g(p) + \#negative\_tags_g(p)\right)} \times 100 \quad (1)$$

After normalization, a high positive score for a certain goal means that statements in the policy related to that goal are mostly on the compliant side; therefore, the service is respectful with that goal. Inversely, a large negative score implies that the policy mostly contains non-compliant statements for that goal; therefore, the service entails serious privacy risks. Finally, a value near 0% means that the number of positive and negative statements for the goal are balanced. Per goal scores give a clear and intuitive picture on how the service provider deals with the different perspectives of data protection.

To assess policies in global, we also quantified the relative difference between positive and negative tags for all goals.

$$Global_{score(p)} = \frac{\sum_{g \in G}(\#positive\_tags_g(p) - \#negative\_tags_g(p))}{\sum_{g \in G}(\#positive\_tags_g(p) + \#negative\_tags_g(p))} \times 100 \quad (2)$$

In this way, it is possible to immediately compare and rank policies of different companies or service providers.

6. Results and Discussion

As introduced above, we trained the SVM with 488 manually tagged sentences. Additionally, we also tested the model by using 366 sentences for training (75%) and the remaining 122 for testing (25%). The average F1 score among the annotation labels considered in the study was 71%. On the one hand, this accuracy is at a similar level to the best accuracy achieved by related works also conducting annotations of privacy policies (albeit with different classes) [8,10]. On the other hand, accuracy is limited by (i) the relatively large number of tags involved in the multi-label classification, which are also not clearly separated; and (ii) the fact that the inter-annotator agreement was not very high, which indicates that the interpretation of the statements in the policies may be subjected to discrepancies even among human experts.

To show the behavior of the system in a real setting, we applied the trained model to both the complete policies from which we extracted the sentences used for training (Google, Mozilla, Netflix and Twitter) and a fresh set of policies of six well-known services and companies: LinkedIn (https://www.linkedin.com/legal/privacy-policy (accessed on 16 February 2021)), Instagram (https://help.instagram.com/519522125107875 (accessed on 16 February 2021)), AliExpress (https://rule.alibaba.com/rule/detail/2034.htm (accessed on 16 February 2021)), Telegram (https://telegram.org/privacy (accessed on 16 February 2021)), Appl (https://www.apple.com/legal/privacy/en-ww/ (accessed on 16 February 2021)) and Huawei(https://www.huawei.com/en/privacy-policy (accessed on 16 February 2021)). When selecting the policies, we avoided picking more than one service offered by the same company (e.g., Instagram/Facebook, YouTube/Google), because the content of their policies may overlap. Moreover, as done for the training data, the fresh set of policies was chosen trying to cover significantly different perspectives regarding privacy protection (generally speaking, Telegram and Instagram have very different insights regarding personal data and privacy).

Figure 1 reports the scores resulting from the automatic multi-label annotation of each policy, both per-goal and in global. Results are sorted from best to worst by their global score.

Intuitively, the ranking seems a good match with the public opinion on how these companies deal with customers' privacy [15]. To conduct a more formal evaluation, we compared our results with the relative ratings provided by Guard (https://useguard.com/ (accessed on 16 February 2021)) for the seven policies that the latter application has already rated, which are, Instagram, Twitter, AliExpress, Netflix, LinkedIn, Mozilla and Telegram. We chose this application because it is the only related work that provides ratings of policies that have been updated according to the legal framework of the GDPR, and that also considers a similar set of services. As mentioned previously, the ratings provided by Guard for each policy consist of a single mark between A+ and F, which aggregates heterogeneous criteria such as generic privacy threats and scandals starred by the company. To quantify the agreement between the two approaches, we measured the correlation between the ranking resulting from sorting policies according to Guard's ratings, and the ranking obtained by sorting the same policies by our global score. Correlation was measured by means of the Spearman's rank correlation coefficient, which assesses how well the relationship between the two scoring criteria can be described using a monotonic function (linear or not). In this way, we avoid depending on the scale of each scoring criteria and we can focus on the relative rankings they define. We obtained a high correlation value

of 0.839, which indicates a strong direct association between the two scores, even though they may consider different criteria.

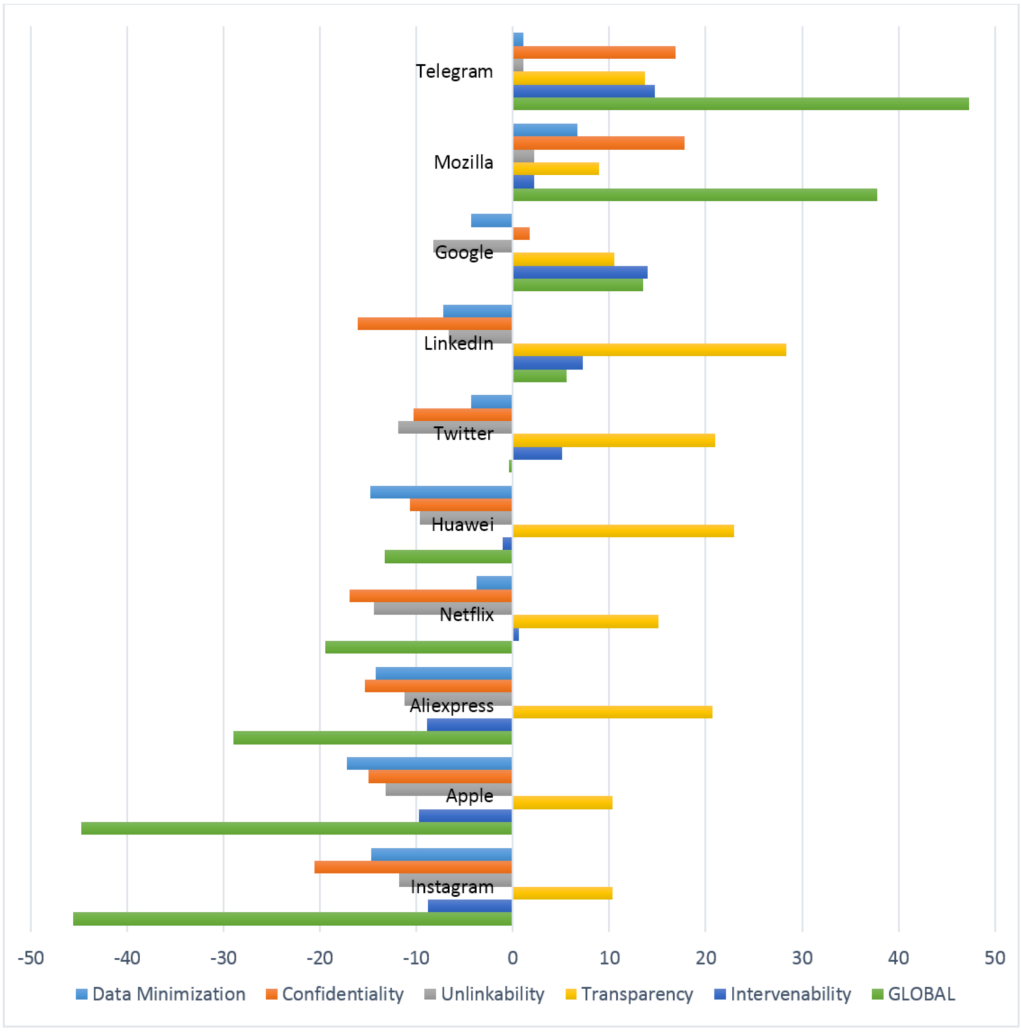

Figure 1. Per-goal and global scores (higher is better) of privacy policies of several services and companies, sorted from best to worst by their global score.

Our approach also provides per-goal scores that help to explain the global score. In particular, these fine-grained scores can make the user aware of the types of risks associated with the use of the service. Focusing on these scores, it can be seen that all service providers fulfill the *transparency* goal. This is expected because the ultimate purpose of a privacy policy is to explain how personal data are managed, regardless how respectful this is with the individual's privacy. In fact, we found a number of sentences that made clear statements (which are good for transparency) on how careless the company is with the customer's data. For instance, Instagram states that it "*cannot ensure the security of any information you transmit to Instagram or guarantee that information on the Service may not be accessed, disclosed, altered, or destroyed*". Being transparent on bad practices indicates that some companies expect customers not to read (or not being able to understand) their policies [3,4].

Regarding the goals that are more closely related to privacy, we can see that the scores associated with *confidentiality*, *unlinkability* and *data minimization* are much more variable across services, but strongly correlated among them for a given service: the average pairwise Pearson correlation among these three goals is 0.811 (with *confidentiality* and *unlikability* exhibiting the strongest correlation of 0.9). The fact that fulfilling these

goals hampers the monetization of customers' data creates a conflict of interest that most companies face by prioritizing their own benefit. In fact, only Mozilla and Telegram, which operate under non-profit principles, obtained positive scores for all these goals. This can be illustrated by comparing how Mozilla and Netflix manage data releases to third parties. Whereas the former state that "*we sometimes release information to make our products better and foster an open web, but when we do so, we will remove your personal information and try to disclose it in a way that minimizes the risk of you being re-identified*", Netflix just informs that "*we disclose your information for certain purposes and to third parties*".

Intervenability obtains middle-of-the-road results. Lacking on *intervenability* goes one step beyond against customers' privacy because it also negates users the possibility of controlling how their data are collected and managed. So, it is not surprising that the services with the worst global score are also the worst regarding *intervenability*: the Pearson correlation between *intervenability* score and the global score is a strong 0.847.

7. Conclusions

We have proposed a system that automatically assesses how respectful a privacy policy is with the data protection goals stated by the GDPR. With this, we aim at making users immediately aware of the privacy threats that are lurking in the shadows of the privacy policies, and allow them to take informed decisions about their rights.

We applied our system to the privacy policies of 10 well-known services and companies. These were selected by taking into account their relevance in society, but also trying to cover significantly different perspectives of privacy protection. The scores provided by our method are sound and consistent with related works. Nevertheless, the real value of our proposal goes far beyond that: we defined clear annotation criteria consistent with the content of the GDPR, and this enabled us not only to provide aggregated scores, but also fine-grained ratings that help to understand the reasons of the assessment. This is aligned with the concept of explainable artificial intelligence, which seeks not only results but clear explanations about how those results have been achieved.

The promising results we report can pave the ground to develop more accurate systems in the near future. In particular, the generality of the classification and its accuracy may be improved by increasing the amount of training data, and by considering not only tech companies, but also other types of services. If large enough training data are available, classifiers other than SVMs may also be employed. Pretrained word embedding models, such as BERT [16], would be especially suitable due to their capacity to be tailored for a specific classification task. Finer grained annotations—e.g., at the level of n-grams rather than sentences—may also contribute to improve the results. Explainability can also be improved by using non-binary tags for each goal. For example, a discrete scale of compliance and non-compliance levels may be defined (and trained) for each goal. The latter, however, will significantly increase the complexity of the training task. We also plan to design and implement a user-friendly interface so that the system can be released and tested by the community on a variety of policies.

Author Contributions: Conceptualization, D.S. and A.V.; data curation, A.V. and M.B.; funding acquisition, D.S.; methodology, D.S. and A.V.; software, D.S.; writing—original draft, D.S.; writing—review and editing, A.V. and M.B. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge support from the European Commission (projects H2020-871042 "SoBigData++" and H2020-101006879 "MobiDataLab"), the Government of Catalonia (ICREA Acadèmia Prize to D. Sánchez and grant 2017 SGR 705), the Spanish Government (projects RTI2018-095094-B-C21 "Consent" and TIN2016-80250-R "Sec-MCloud") and the Norwegian Research Council (project no. 308904 "CLEANUP").

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors are with the UNESCO Chair in Data Privacy, but the views in this paper are their own and are not necessarily shared by UNESCO.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. FaceApp's Terms of Use. Available online: https://www.faceapp.com/terms-20170803.html (accessed on 23 December 2020).
2. Koetsier, J. Viral app FaceApp Now Owns Access to More than 150 Million People's Faces and Names, in Forbes, ed. 17 July. Available online: https://www.forbes.com/sites/johnkoetsier/2019/07/17/viral-app-faceapp-now-owns-access-to-more-than-150-million-peoples-faces-and-names/ (accessed on 16 February 2021).
3. Obar, J.A.; Oeldorf-Hirsch, A. The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services. *Inf. Commun. Soc.* **2018**, *23*, 128–147. [CrossRef]
4. Litman-Navarro, K. We Read 150 Privacy Policies. They Were an Incomprehensible Disaster, in New York Times, ed. 12 June. Available online: https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html (accessed on 16 February 2021).
5. Zimmeck, S.; Bellovin, S.M. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, 20–22 August 2014; pp. 1–16.
6. Wilson, S. The Creation and Analysis of a Website Privacy Policy Corpus. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 1330–1340.
7. Harkous, H.; Fawaz, K.; Lebret, R.; Schaub, F.; Shin, K.G.; Aberer, K. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In Proceedings of the 27th USENIX Security Symposium, Balimore, MD, USA, 15–17 August 2018; pp. 531–548.
8. Zaeem, R.N.; German, R.L.; Barber, K.S. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Trans. Internet Technol.* **2018**, *18*, 39. [CrossRef]
9. Tesfay, W.B.; Hofmann, P.; Nakamura, T.; Kiyomoto, S.; Serna, J. PrivacyGuide: Towards and Implementation of the EU GDPR on Internet Privacy Policy Evaluation. In Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics, Tempe, AZ, USA, 21 March 2018; pp. 15–21.
10. Müller, N.M.; Kowatsch, D.; Debus, P.; Mirdita, D.; Böttinger, K. On GDPR Compliance of Companies' Privacy Policies. In Proceedings of the 22nd International Conference on Text, Speech, Dialogue, Liubliana, Eslovenia, 10–13 September 2019; pp. 151–159.
11. Conference of the Independent Data Protection Supervisory Authorities of the Federation and the Länder, The Standard Data Protection Model Version 2.0b. Available online: https://www.datenschutzzentrum.de/uploads/sdm/SDM-Methodology_V2.0b.pdf (accessed on 16 February 2021).
12. Feng, S.; Zhou, H.; Dong, H. Using deep neural network with small dataset to predict material defects. *Mater. Des.* **2019**, *162*, 300–310. [CrossRef]
13. Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. A Practical Guide to Support Vector Classification, National Taiwan University. 2016. Available online: https://www.csie.ntu.edu.tw/~{}cjlin/papers/guide/guide.pdf (accessed on 16 February 2021).
14. Lebanoff, L.; Liu, F. Automatic Detection of Vague Words and Sentences in Privacy Policies. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3508–3517.
15. Newton, C. The Verge Tech Survey 2020, in The Verge, ed. 2 March. Available online: https://www.theverge.com/2020/3/2/21144680/verge-tech-survey-2020-trust-privacy-security-facebook-amazon-google-apple (accessed on 16 February 2021).
16. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2018. Available online: https://arxiv.org/abs/1810.04805v2 (accessed on 16 February 2021).