Detection and nudge-intervention on sensitive information in social networks

J. ALEMANY[*], *VRAIN, Universitat Politècnica de València, Camino de Vera s/n, 46021 Valencia, Spain.*

V. BOTTI-CEBRIÁ, *VRAIN, Universitat Politècnica de València, Camino de Vera s/n, 46021 Valencia, Spain.*

E. DEL VAL, *Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Escuela Universitaria Politècnica de Teruel, 44003 Teruel, Spain.*

A. GARCÍA-FORNES, *VRAIN, Universitat Politècnica de València, Camino de Vera s/n, 46021 Valencia, Spain.*

Abstract

Detecting sensitive information considering privacy is a relevant issue on Online Social Networks (OSNs). It is often difficult for users to manage the privacy associated with their posts on social networks taking into account all the possible consequences. The aim of this work is to provide information about the sensitivity of the content of a publication when a user is going to share it in OSN. For this purpose, we developed a privacy-assistant agent that detects sensitive information. Based on this information, the agent provides a message through a nudge mechanism warning about the possible risks of sharing the message. To avoid being annoying, the agent also considers the user's previous behaviour (e.g. if he previously ignored certain nudges) and adapts the messages it sends to give more relevance to those categories that are more important to the user from the point of view of the privacy risk. This agent was integrated into the social network PESEDIA. We analysed the performance of different models to detect a set of sensitive categories (i.e. location, medical, drug/alcohol, emotion, personal attacks, stereotyping, family and association details, personal details and personally identifiable information) in a dataset of tweets in Spanish. The model that obtained the best results (i.e. F1 and accuracy) and that was finally integrated into the privacy-assistant agent was transformer-based.

Keywords: Privacy, information sensitivity, social networks, classification.

1 Introduction

Online social networks (OSNs) have become a pillar of modern society and have changed the way people communicate with each other [21]. Their usage provides users with benefits, such as entertainment, influencing others, receiving support, maintaining relationships (even creating new ones) and increasing their reputation, in exchange for relinquishing some privacy. The way to control this exchange between social benefit and privacy loss is through privacy policies. Although some social networks provide privacy mechanisms to protect users, these can be complex to use and configure, so users may keep the default settings, as they do not know the risk behind what they are trying to post. Moreover, posts may contain personal information of different types and levels of

[*] E-mail: jalemany1@dsic.upv.es

Vol. 30, No. 6, © The Author(s) 2022. Published by Oxford University Press. All rights reserved.
For permissions, please e-mail: journals.permission@oup.com.
Advance Access published 21 February 2022 https://doi.org/10.1093/jigpal/jzac004

sensitivity such as private life events, sexual preferences, diseases or political ideas. As a result, the users' information may be accessed by unknown people or companies for non-benevolent purposes (i.e. phishing, bullying, stalking, marketing campaigns and commercial usage).

To deal with this problem, research has made advances in the analysis and detection of properties in text and images from social networks (such as sentiment analysis, hate speech detection and private information detection) [2, 8, 9, 26]. Also, campaigns have been carried out to make people aware of the implications of sharing their data on social networks [15]. Nevertheless, some studies consider that awareness and trust do not necessarily promote less risky behaviour, especially among young people. This result is in line with the number of young people who report negative experiences despite the initiatives carried out by educational campaigns [14]. As an alternative to educational materials and campaigns, it has been considered that tools or mechanisms integrated into social networks that assist users in making better privacy decisions can reduce their exposure to privacy risks [3]. Specifically, soft-paternalism interventions have been considered as an appropriate method to influence users' privacy behaviours, without losing their freedom, towards less risky actions from the point of view of privacy [6].

In this paper, we aim to deal with the problem of sharing text publications containing sensitive information in social media using techniques based on the idea of soft-paternalism to help users to understand what kind of information are they giving to other people. To do this, we propose a privacy-assistant agent that analyses the content of the publication, detects if there is information belonging to potentially sensitive categories and advises the user to help him in the decision-making process. Therefore, the main contributions of this work are the following: (i) a privacy-assistant agent that assesses the sensitivity of users' posts based on a set of sensitive categories, (ii) the detection of sensitive information in Spanish text messages, (iii) the adaptation of the agent's behaviour to users' perception and previous behaviours (i.e. their risk acceptance), (iv) the generation of informational messages during privacy decisions to aid the user to make informed decisions, (v) the integration of these contributions (via the privacy-assistant agent) into the social network PESEDIA[1] and (vi) the development of a dataset of tweets in Spanish to analyse the performance of the classifiers to detect the categories established.

The paper is organized as follows. Section 2 presents previous works related to sensitive information and automatic detection of sensitive categories. Section 3 describes the privacy-assistant agent for detecting information sensitivity. Section 4 evaluates the proposal through a set of experiments using a Twitter dataset. Finally, Section 5 presents some conclusions and final remarks.

2 Related work

An important factor of online disclosure actions in social networks is the information that users share. This information may contain personal data, which is information relating to an individual that can be used to identify that individual directly or indirectly and varies in types and levels of sensitivity. The sensitivity of the information can be analysed from the point of view of intimacy and is associated with the potential loss of privacy. The greater intimacy of the information, the riskier and more uncomfortable its disclosure [19]. However, on social networks, users perceive benefits as close while risks as abstract and psychologically distant [16]. Moreover, users are often unaware of the sensitivity of the information or they have different perceptions of sensitivity. For example, religion is a highly sensitive topic in areas where there is a high degree of sectarian conflict but of a

[1] https://pesedia.webs.upv.es/

low degree in other areas. As a result, the online disclosure actions may not be free of regret and so users may face potential negative effects. These negative effects would include psychological (e.g. loss of self-concept due to embarrassment), physical (e.g. loss of life or health) or material (e.g. loss of financial or other assets) consequences [22]. For this reason, it is important to understand the potential risks that a user may be exposed to when sharing sensitive information on a social network.

Previous works present proposals to address the problem of loss of privacy when sharing information on social networks. Most of them check users' privacy settings for metrics and scores. Alemany *et al.* [4, 5] propose a privacy risk metric that estimates the reachability of a user's sharing action based on the distance between the user and the potential audience that might see the publication. However, the value of this metric lacks an assessment of the sensitivity of the information that users share in social networks. Pensa and Di Blasi [20] also propose a theoretical framework to measure the privacy risk based on the sensitivity and visibility of a profile item i published by user u. This measure is used to assists user u to personalize their privacy settings. Talukder *et al.* [25] propose a privacy protection tool that measures the amount of user's profile sensitive information that can be inferred from the profile of user's friends. Based on the value of leakage, the tool also recommends actions to reduce the amount of leakage asking friends to hide certain types of information. However, these last two approaches are focused on the sensitivity of profile information and do not take into account the information shared in posts.

To deal with the sensitivity analysis of users' posts, we have identified in the literature two approaches: (i) one based on semantics, Information Content theory, and words' taxonomy; and (ii) another based on Natural Language Processing (NLP) and Machine Learning techniques. In the first approach, we highlight the work presented by Sanchez *et al.* [23]. They propose an automatic method to assess the sensitivity of textual publications considering the user's privacy requirements towards other users in the social network. The level of sensitivity is established based on the semantics of the terms contained in the publication. They consider that the terms that contain more information are the ones that disclose more knowledge to attackers. To define the user's privacy requirements associated with a level of sensitivity, the user is asked to define the maximum knowledge that would be disclosed to each privacy level/type of relationship with other users. However, it lacks weighted categories that specify the maximum and minimum sensitivity value for each type of information, and, to the best of our knowledge, this work is not integrated into a social network. In the second approach (and most popular), we find the following works. One of the first works in analysing the content that users usually leak in social networks was presented by Mao *et al.* [18]. They specifically focus their analysis on three types of leaks: disclosing vacation plans, tweeting under the influence of alcohol and revealing medical conditions. Based on this information, they build classifiers to automatically detect these three types of leaks on tweets. However, this approach is limited to a reduced set of sensitive categories. Another recent project [27] classified private information into 13 different (potentially) sensitive categories, using the common TF-IDF, Bag-of-Words and sentiment methods. However, they used simple, supervised classifiers such as the Naive Bayes classifier, which cannot capture semantic features or accurately discover categories containing subtle, yet sensitive, content. By using deep learning models, Wang *et al.* [28] propose a context-aware, text-based quantitative model for private information assessment, namely PrivScore, which serves as the foundation of a privacy leakage alerting mechanism. They examine the responses collected on the sensitivity of private information from crowd-sourcing workers' opinions. They discover a perceptual model behind the consensuses and disagreements by using deep neural networks. The way they assess the sensitivity of information is by labelling information into five categories: non-sensitive, maybe, little sensitive, sensitive and very sensitive. However, a broader evaluation scale should be used to more accurately capture the sensitivity of the information. Moreover, these advances may become outdated due to the latest and promising advances made in NLP with the usage of transformer-based pre-trained models [17].

As we stated, assessing the privacy risk (e.g. from the sensitivity of the information) is not enough, this information should be provided to users during privacy decisions. The aim is to promote behaviours towards more secure privacy policies. In this line, there are interesting works that propose the intervention via alerts and advice. For instance, Wang *et al.* perform several experiments where nudges are integrated into the Facebook social network [29, 30] to advise about privacy risks. One of the nudges provides images of the audience that could see the post and the potential audience in the case of re-sharing actions of the publication. Other nudges are oriented to think twice before posting content on a social network. The authors propose a 'timer nudge', which includes a time delay before the user posts a message on the social network, and a 'sentiment nudge', which consists of an estimation of the sentiment associated with the post that the user is going to publish. Alemany *et al.* [3] also propose the use of nudges integrated into a social network. These nudges inform about the degree of privacy risk of publishing content for a specific audience. The user can assess whether the audience selected to view the publication is appropriate or not. The results of the experiments performed provide evidence of changes in posting behaviour for some of the participants. Another approach has been done by Wang *et al.* [28]. In this work, they collect and classify their dataset, create a sensitivity measure and detect certain types of information and create a sensitivity indicator to measure how much sensitive information users are publishing. This tool is appropriate to summarize the general users' behaviour regarding privacy in social networks but not to notice/inform users in specific moments of privacy decisions.

In this work, we propose a privacy-assistant agent to compute the sensitivity of the information and to provide users with this information (using nudge mechanisms) before they publish it. This agent takes into account a set of the most common-sensitive information categories (e.g. location, personal data, personal attacks, etc.) to identify personal, sensitive information in users' posts. To detect them, we analyse the performance of several techniques (such as machine learning, entity recognition, ontologies, dictionaries, sentiment analysis and hate speech detection) on a dataset of labelled Twitter posts.

3 The privacy-assistant agent

In order to help users make informed decisions about whether or not to post a text on a social network, we propose a privacy-assistant agent that assesses the sensitivity of the information of a post and gives feedback to users before they share it. To do this, we first define what we consider to be sensitive information and which categories have been taken into account to represent sensitivity. We then describe the process followed by the privacy-assistant agent to detect these categories in a post. Finally, we illustrate the agent's interaction with the user via an informed nudge message displayed at the time of privacy decision making. This privacy-assistant agent is integrated into the social network Pesedia.

According to the works analysed in the literature and the current regulations (European General Data Protection Regulation (GDPR)) [1], we have considered defining a set of the most common-sensitive information categories to simplify the task of automatic detection of sensitive content. In the case of the GDPR, it defines a set of categories of data considered as sensitive where the following categories would be: racial origin, political opinions or religious or other beliefs as well as personal data on health, sex life or criminal convictions. In the case of the literature's proposals, we find interesting the one proposed by Caliskan *et al.* [12]. They detect ten relevant categories from the point

TABLE 1. Sensitive information categories and its meaning.

Category	Description
Location	The content of the post discloses information relating to both a specific location (e.g. a city or address) and a partial location (e.g. at the cinema or at home).
Medical	The content of the post reveals information about someone’s medical condition, from diseases to symptoms or restrictions.
Drug/alcohol	The content of the post gives information about drug/alcohol use or discloses information under its influence.
Emotion	The content of the post is highly emotional, euphoria, frustration, hot states, etc.
Personal attacks	The content of the post contains critical statements directed at an individual, also including the use of insults and foul language.
Stereotyping	The content of the post contains critical and/or stereotypical statements directed at collectives/groups (e.g. ethnic, racial, national, religious, political).
Family/association details	The content of the post reveals information about family members, or reveals their associations (e.g. ex-partner, mother-in-law, step brother, employee).
Personal details	The content of the post reveals personal details (e.g. relationship status, sexual orientation, beliefs, job/occupation, embarrassing or inappropriate content, reveal/explain too much).
Personally identifiable information	The content of the post contains personally identifiable information (e.g. credit card number, email address, phone number, home address, birth date).
Neutral/objective	The content of the post is neutral or objective, that is, it does not reveal private or sensitive information.

of view of sensitivity through an analysis of Twitter publications. In these categories, you will find location, medical, drug/alcohol, emotion, personal attacks, stereotyping, family/association detail, personal details, personally identifiable information and neutral/objective information. Moreover, they distinguish whether the information comes from the author of the post or it refers to another user. Our proposal combines Caliskan’s work categories with the sensitive information categories defined in the GDPR. Table 1 presents the categories of sensitivity taken into account in this work and their meaning.

The privacy-assistant agent is responsible for gathering the writing of users in the text field of the social network and providing users with informed nudges about the sensitivity of the information before they share the post (i.e. click ‘Publish’). Specifically, the agent gathers a text each time the user stops writing for a few seconds (around 2 seconds). The agent then applies a pre-processing text phase to clean the input by removing links, mentions, and unusual characters. The agent also takes into account the emoticons and emojis (does not remove them), because they may contain valuable information for the detection of some of the sensitive categories. Depending on the models used to predict each one of the sensitive categories in a text message, it is applied some other pre-processing functions to prepare input for the models (e.g. remove stop-words, tokenization, etc.). We

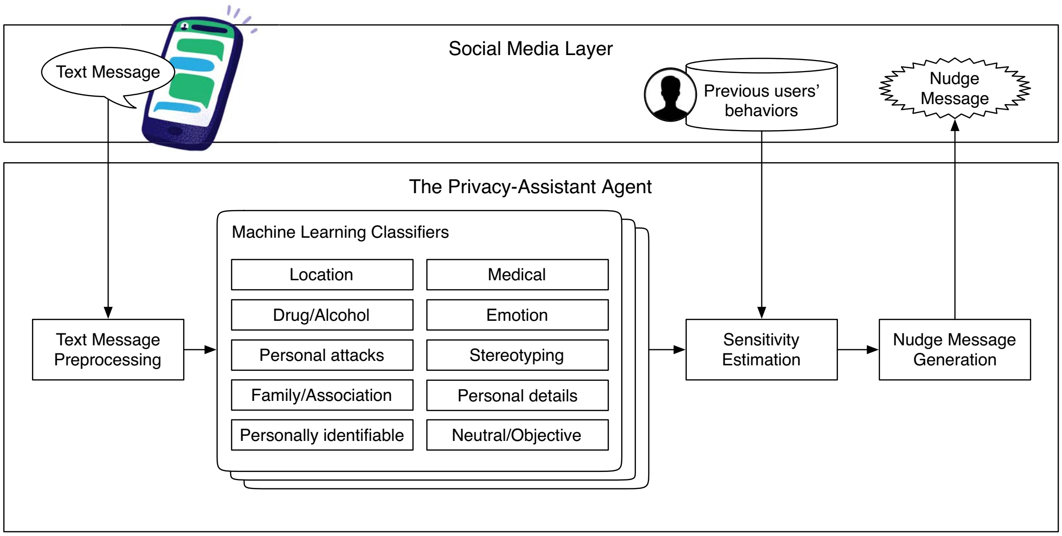


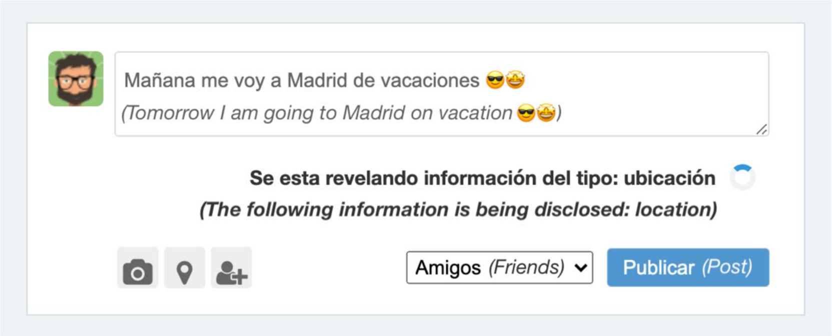
FIGURE 1. The privacy-assistant agent workflow.

have trained a model for detecting each one of the sensitive categories considered in this work (see Table 1). Once the models' output notices that a text may contain some sensitive category, the agent processes all these results and combine them with the previous user's behaviours. Thus, the agent provides more relevance to sensitive categories that users do not usually disclose than the ones users regularly do. Algorithm 1 explains the process that the agent follows to weigh the relevance of each sensitive category for a user. The relevance of a sensitive category follows a decreasing logarithmic function, decreasing the relevance each time users accept sharing information of this category and increasing it with the opposite action and overtime. Finally, the agent generates a nudge message warning the user about the sensitive information contained in the user's post. Figure 1 shows the different phases through which the agent goes through (i.e. collection, pre-processing, detection and interaction with the user). This process can be repeated as many times as users need to finally post the text message.

Algorithm 1 Compute relevance of a sensitive category for a user.

```
procedure RELEVANCEOF($C$, $U$):
  $userRiskAccepts$ = get_posts(owner=$U$, category=$C$, count=True)
  $userLastAccept$ = get_last_post_timestamp($U$, $C$)
  $value = log_2\left(\frac{\beta \cdot userLastAccept}{\alpha \cdot (1+userRiskAccepts)}\right)$
  if $value > 0$ then
    Return $True$
  else
    Return $False$
  end if
end procedure
```

Note: α and β are hyper-parameters used to ponder variables in the logarithmic function.



Mañana me voy a Madrid de vacaciones 😎🤩
(Tomorrow I am going to Madrid on vacation 😎🤩)

Se esta revelando información del tipo: ubicación
(The following information is being disclosed: location)

FIGURE 2. Screenshot of the posting field in PESEDIA with an example of a text message and the response of the privacy-assistant agent (the informative nudge message). Note: text's translations to English were included for the paper.

The privacy-assistant agent was fully integrated into a real social network called PESEDIA. PESEDIA is an online social network for educational and research purposes developed by the Valencian Research Institute for Artificial Intelligence (VRAIN). This social network is developed in Elgg [13], an open-source engine used mainly for the creation of social network environments. The environment provided by this engine is similar to other social networks (e.g. Facebook). We developed each functionality of the agent in PESEDIA through modules following the design principles of the Elgg engine. The modules allow us to enable and disable online features of the social network at any time, adapting them to the needs of the experiment. Figure 2 shows the final integration of the privacy-assistant agent in the social network and its interaction with the users' daily posting activities.

4 Experiments

In this section, we evaluate the performance of the decision process followed by the privacy-assistant agent to detect the different sensitive categories. For the evaluation, we created a dataset of tweets labelled with the categories considered as sensitive. To measure the performance of the models, we have evaluated them using metrics such as accuracy and macro F1. The use of the macro version is imperative to avoid any possible misleading result caused by huge class imbalance scenarios.

4.1 Dataset

To build the dataset, we collected posts from Twitter because of its popularity and ease of access. We performed a snowball crawling process and collected 785,403 tweets from 1,697 Spanish users. We eliminated non-Spanish tweets (about 189,233) and then selected the potentially sensitive tweets by filtering them with keywords extracted from Wang's work [28] and translated to the Spanish language. A total of 10,683 tweet posts were selected for the labelling task. Annotators were asked to annotate each tweet as many times as sensitive categories, i.e. location, medical, drug/alcohol, emotion, personal attacks, stereotyping, family/association details, personal details, personally identifiable information and neutral/objective information (see Table 1); marking as true (if it contains) or false (if it does not contain) each of these categories. Therefore, a tweet may belong to several categories. Moreover, each tweet was annotated by four annotators previously trained in

TABLE 2. Number of tweets per sensitive category and inter-rater agreement values.

Category	# tweets	Fleiss Kappa	PABAK
Location	247	0.61	0.93
Medical	144	0.76	0.97
Drug/alcohol	73	0.79	0.99
Emotion	1,600	0.49	0.56
Personal attacks	381	0.60	0.89
Stereotyping	133	0.58	0.96
Family/association details	392	0.59	0.89
Personal details	421	0.43	0.85
Personally identifiable information	32	0.70	0.99
Neutral/objective	1,801	0.51	0.55

the detection of these sensitive categories. Finally, annotators were able to annotate a set of 3,707 tweets.

To examine the consistency among annotators, the quality of the labelling was analysed. We assess the inter-rater agreement using Fleiss Kappa [15], which values ranges from 0 (poor agreement) to 1 (perfect agreement). However, considering only this statistic is not appropriate when the prevalence of a given response is very high or very low in a specific class [7]. In order to address these imbalances caused by differences in prevalence and bias, we also assess the PABAK coefficient [10], which depends solely on the observed proportion of agreement among annotators. The PABAK coefficient values range from –1 to 1, with 0 being 50% agreement among the annotators. Table 2 shows the number of tweets finally stated for each of the sensitive categories and the results of the inter-rater agreement.

It can be seen that the dataset is imbalanced because there are quite differences in the number of samples among classes. Nevertheless, the level of agreement of the annotators is still quite good for most of the sensitive categories. Only in a few categories such as emotion, personal details and neutral/objective information, there was a fair/moderate agreement among the annotators. Also, as a result of the annotation, there were tweets with no labels (132) because no consensus among annotators was reached. Therefore, the resultant dataset is composed of 3,575 tweet posts.

4.2 Evaluation

For the task of automatic detection of sensitive content in OSN posts, we have trained different classification models to evaluate the performance of each one in the sensitive categories detection. The best classification models will be used by the privacy-assistant agent to aid users in the privacy decision-making process. In the case of machine learning models, we have used the *scikit-learn*[2] library, which provides a large number of classifiers and different techniques to process the text to be classified. We considered the following models: Random Forest (estimators = 1,000, random state = 0), Naïve Bayes (Default), Linear Support Vector Machine (Default) and KNN (2 neighbours), being these models the most effective in text categorization [24]. In the case of deep learning models, we

[2] https://scikit-learn.org/stable/

have used the *pytorch*[3] and *transformers*[4] libraries. By using these libraries, we applied Inductive Transfer Learning combined with the BETO transformer pre-training method (a Spanish variant of BERT transformer) [11] that allow us to learn our task, not from scratch but using previously calculated weights. Moreover, the text tokenization using transformers is context-dependent with a really large vocabulary (about 30,000 entries), which allows a better representation of inputs for the deep learning models. Specifically, we trained BETO with a maximum sequence length of 128, a batch size of 32, 20 epochs and a learning rate of 2e-5.

One of the problems of the dataset used is that most of the categories are imbalanced. To deal with this problem, we tested the following methods: upsampling and downsampling techniques, and also a weighted sample loader in the case of the deep learning model. However, no significant results were obtained in comparison with the original sampling, except by the downsampling technique that slightly improves the model results. Moreover, we performed 10-fold cross-validation, which splits the samples into sets of 10% for test and 90% for training, using the stratified version that is recommended for imbalanced problems because it maintains the ratio between the target classes.

To assess the performance of trained models, we use the metrics of accuracy and macro F1. Accuracy provides us information about the fraction of predictions our model got right indirectly giving more importance to the majority class, while macro F1-score provides us a balance between precision and recall but giving the same importance to each class. In binary problems like in this work's task, macro F1-score will be low for models that only perform well on the majority class while performing poorly on the minority class (i.e. values lower than 50/100). Table 3 contains the average results of cross-validation obtained with the different trained models. The performance obtained by the BETO (transformer-based) model stands out over the other types of models with quite good accuracy values (greater than 90%) and macro F1-scores ranging from 70% to 80% for location, medical, drug/alcohol, emotion, personal attacks and family/association details categories. Only in the personal details and personally identifiable information categories has been observed more poorly results. For the category of personally identifiable information, these results may easily be caused by the low number of samples of this class (32). For the category of personal details, these results are the poorest, which matches the low agreement among annotators for this class. The other models also obtained good results in some tasks, mainly SVM models in drug/alcohol, emotion and neutral/objective categories and Random Forest models in medical and personal attacks categories; however, they do not obtain better results than the BETO model.

The final privacy-assistant agent implementation includes the transformer-based models. Our agent proposal has been used in the social network Pesedia, where approximately 200 users have been using the network while the agent analysed the sensitive information of each of the texts that users intended to post. The average time needed by the privacy-assistant agent to process a text and give feedback is very small (about 0.27 seconds), so users should not wait for practically anything to obtain the sensitivity analysis. A point of improvement would be to replicate the privacy-assistant agent to attend simultaneously several requests. These results support the robustness of the agent proposal, which real-time responses can lead users to more informed privacy decisions.

5 Conclusions

In this paper, we have proposed a privacy-assistant agent for the assessment of sensitive information in OSN publications. The agent provides information to the user about which sensitive categories

[3] https://pytorch.org

[4] https://huggingface.co/transformers/

TABLE 3. Performance of the models on automatic detection of sensitive categories, given in macro F1 and accuracy. Note: BETO, Transformer-based Inductive Transfer Learning; RF, Random Forest; NB, Naive Bayes; SVM, Support Vector Machine; K-NN, K-Nearest Neighbours.

	BETO		RF		NB		SVM		K-NN	
Category	F1	(Acc.)	F1	(Acc.)	F1	(Acc.)	F1	(Acc.)	F1	(Acc.)
Location	**72.6**	**(94.4)**	54.1	(92.8)	47.7	(68.3)	53.8	(93.3)	50.1	(93.2)
Medical	**87.2**	**(98.1)**	71.6	(96.8)	47.9	(75.6)	69.5	(96.7)	52.0	(96.1)
Drug/alcohol	**80.53**	**(99.0)**	67.9	(98.1)	49.4	(84.8)	74.2	(98.7)	52.2	(98.0)
Emotion	**69.4**	**(69.9)**	61.1	(62.2)	51.8	(54.4)	62.2	(63.5)	44.4	(56.2)
Personal attacks	**69.4**	**(91.0)**	57.2	(87.2)	45.5	(54.1)	54.7	(88.0)	48.6	(87.4)
Stereotyping	**66.9**	**(96.8)**	51.4	(96.5)	45.7	(73.2)	50.8	(96.4)	49.6	(96.4)
Family/association details	**76.7**	**(92.2)**	54.6	(88.3)	52.8	(65.8)	54.8	(89.6)	54.2	(88.9)
Personal details	**48.7**	**(88.3)**	48.4	(86.5)	42.5	(51.7)	46.9	(88.2)	46.8	(87.7)
Personally identifiable information	**64.8**	**(99.3)**	55.0	(99.1)	47.2	(86.5)	55.0	(99.1)	55.0	(99.1)
Neutral/objective	**65.1**	**(65.2)**	61.0	(62.2)	51.8	(54.4)	62.2	(63.5)	44.4	(56.2)

have been detected to facilitate the decision process of posting or not a message. The agent also considers the previous behaviours of the user to personalize the messages considering the user's perception of the sensitivity of certain categories and avoiding being annoying with the messages. We have evaluated different models for the detection of sensitive categories considering a dataset of tweets in Spanish. The results show that the transformer-based model (i.e. BETO model) offers better results in all the sensitive categories considered than the commonly used classifiers. Only in the personal details and personally identifiable information categories, the transformer-based model provided more poorly results. These results could be caused by the low number of samples or the low agreement among annotators for these categories. We plan to extend the dataset to balance the samples in the categories. We will consider the combination of an audience estimator with the content sensitivity analysis to provide a more complete view of the privacy risk.

Acknowledgements

This work is supported by the Spanish Government project TIN2017-89156-R.

References

[1] Official legal text

[2] G. Aguado, V. Julian and A. Garcia-Fornes. Towards aiding decision-making in social networks by using sentiment and stress combined analysis. *Information*, **9**, 107, 2018.

[3] J. Alemany, E. del Val, J. Alberola and A. García-Fornes. Enhancing the privacy risk awareness of teenagers in online social networks through soft-paternalism mechanisms. *International Journal of Human-Computer Studies*, **129**, 27–40, 2019.

[4] J. Alemany, E. del Val, J. Alberola and A. García-Fornes. Estimation of privacy risk through centrality metrics. *Future Generation Computer Systems*, **82**, 63–76, 2018.

[5] J. Alemany, E. del Val, J. M. Alberola and A. Garćia-Fornes. Metrics for privacy assessment when sharing information in online social networks. In *IEEE Access*, IEEE, 2019.

[6] J. Alemany, E. del Val and A. García-Fornes. Empowering users regarding the sensitivity of their data in social networks through nudge mechanisms. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, pp. 2539–2548, January 2020. University of Hawaii at Manoa, Association for Information Systems IEEE Computer Society Press.

[7] M. Anzovino, E. Fersini and P. Rosso. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pp. 57–64. Springer, 2018.

[8] A. Bisht, H. S. Annapurna Singh and J. V. Bhadauria., *et al.* Detection of hate speech and offensive language in twitter data using lstm model. In *Recent Trends in Image and Signal Processing in Computer Vision*, pp. 243–264. Springer, 2020.

[9] V. Botti-Cebriá, E. del Val and A. García-Fornes. Automatic detection of sensitive information in educative social networks. In *Conference on Complex, Intelligent, and Software Intensive Systems*, pp. 184–194. Springer, 2020.

[10] T. Byrt, J. Bishop and J. B. Carlin. Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, **46**, 423–429, 06 1993.

[11] J. Cañete, G. Chaperon, R. Fuentes and J. Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020* (to appear), 2020.

[12] A. C. Islam, J. Walsh and R. Greenstadt. Privacy detective: detecting private information and collective privacy behavior in a large social network. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*. ACM, 2014.

[13] C. Costello. Elgg 1.8 Social Networking. Packt Publishing Ltd, 2012.

[14] A. Dhir, P. Kaur, S. Chen and K. Lonka. Understanding online regret experience in facebook use–effects of brand participation, accessibility & problematic use. *Computers in Human Behavior*, **59**, 420–430, 2016.

[15] R. Falotico and P. Quatto. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, **49**, 463–470, 2015.

[16] C. Hallam and G. Zanella. Online self-disclosure: the privacy paradox explained as a temporally discounted balance between concerns and rewards. *Computers in Human Behavior*, **68**, 217–227, 2017.

[17] J. Lin, R. Nogueira and A. Yates. Pretrained transformers for text ranking: BERT and beyond. Preprint, arXiv:2010.06467, 2020.

[18] H. Mao, X. Shuai and A. Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society*, pp. 1–12. ACM, 2011.

[19] D. L. Mothersbaugh, W. K. Foxx, S. E. Beatty and S. Wang. Disclosure antecedents in an online service context: the role of sensitivity of information. *Journal of Service Research*, **15**, 76–98, 2012.

[20] R. G. Pensa and G. Di Blasi. A privacy self-assessment framework for online social networks. *Expert Systems with Applications*, **86**, 18–31, 2017.

[21] F. Requena and L. Ayuso. Individualism or complementarity? The effect of digital personal networks on face-to-face personal networks. *Information, Communication & Society*, **22**, 2097–2111, 2019.

[22] J. M. M. Rumbold and B. K. Pierscionek. What are data? A categorization of the data sensitivity spectrum. *Big Data Research*, **12**, 49–59, 2018.

[23] D. Sánchez and A. Viejo. Privacy risk assessment of textual publications in social networks. In *ICAART*, pp. 236–241, Science and Technology Publications, Lda., 2015.

[24] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, **34**, 1–47, 2002.

[25] N. Talukder, M. Ouzzani, A. K. Elmagarmid, H. Elmeleegy and M. Yakout. Privometer: privacy protection in social networks. In *The 2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*. IEEE, 2010.

[26] J. Taverner, R. Ruiz, E. del Val, C. Diez and J. Alemany., eds. Image analysis for privacy assessment in social networks. In *International Symposium on Distributed Computing and Artificial Intelligence*, pp. 1–4. Springer, 2018.

[27] Q. Wang, J. Bhandal, S. Huang and B. Luo. Content-based classification of sensitive tweets. *International Journal of Semantic Computing*, **11**, 541–562, 2017.

[28] Q. Wang, H. Xue, F. Li, D. Lee and B. Luo. #DontTweetThis: scoring private information in social networks. In *Proceedings on Privacy Enhancing Technologies*, **2019**, 72–92, 2019.

[29] W. Yang, P. G. Leon, A. Acquisti, L. F. Cranor, A. Forget and N. Sadeh. A field trial of privacy nudges for facebook. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pp. 2367–2376. ACM, 2014.

[30] Y. Wang, P. G. Leon, X. Chen and S. Komanduri. From Facebook regrets to Facebook privacy nudges. *Ohio St. LJ*, **74**, 1307, 2013.

Received 20 February 2021