Sensitive Content Recognition in Social Interaction Messages

Isidoros Perikos [a, 1] and Loizos Michael [a,b]
[a] *Open University of Cyprus, Nicosia, Cyprus*
[b] *CYENS Center of Excellence, Nicosia, Cyprus*

Abstract. Online social networks are a predominant medium for social interaction where people communicate in a way similar to what they do in real life. User communication comes mainly in forms of textual data which are rich in personal information, opinions and sentiments. The automatic recognition of sensitive content in texts in online social networks is quite important for a number of reasons. In this work, we study the dimensions of sensitive content recognition and we examine the performance of various machine learning methods for sensitive data recognition in text. Understanding the key features of sensitive content can assist in formulating more efficient user-centric interaction frameworks too that secure users' privacy, promote users' inclusion and enhance the diversity awareness of the online society. Also, another part of this work focuses on the models' explainability where the integration of LIME and SHAP offer insight on features that are consistent and robust predictors of sensitive content.

Keywords. social networks, user interactions, sensitive content, explainability

1. Introduction

The proliferation of social networks has increased our capacity to interact, communicate, and network [1], by creating new online environments to facilitate user interactions [2], and do so in a way similar to what users do in real life [5], [6]. User interactions come mainly through the exchange of textual data, rich in personal information, opinions, and sentiments. The automatic recognition of sensitive content in user interactions becomes, thus, critical. Methods that automatically identify sensitive data can facilitate smoother user interactions, and can assist users to be freely involved in online interactions and communications, by protecting, for example, minorities and marginalized groups from being attacked by others. In addition, the detection of sensitive data can assist in facing hate speech and discrimination too.

In this paper, we present machine learning models that were trained to detect sensitive content and examine their performance under different case scenarios. The dataset used for training and testing includes real-life user-generated data that were gathered during a pilot study of the WeNet platform, and they were annotated in terms of their sensitive nature by an Ethics expert. Typical post-hoc explainability techniques were also used to offer insights on what parts of each data-point contribute to its sensitive nature, allowing us to identify words that are consistent and robust predictors of sensitivity across our dataset, as well as rare keywords that can instantly swing a prediction towards between being sensitive or not.

[1] Author emails: {isidoros.perikos, loizos}@ouc.ac.cy

2. Related Work

Over the last decade, there is a huge research interest in sensitive content detection methods. An overview of methods and approaches can be found in [7], [12]. Authors in [8], introduce an approach for automatically identifying sensitive information from text in a manner that is not specific to any particular domain. The method utilizes information theory and leverages a large corpus to evaluate the sensitivity level of terms based on the information they convey with quite good performance. In [9], a scheme for detecting sensitive information in unstructured text using a Text-CNN approach is presented. By leveraging Text-CNN and analyzing the context and semantics of sentences, this scheme avoids inaccuracies caused by manually defining sensitive words. It also enables the establishment of the detection model according to real-world requirements, resulting in an efficient security detection process. In [10] authors investigate the suitability of pre-trained transformer models for detecting complex sensitive information. Through experiments conducted on the Monsanto trial dataset, authors observed that the fine-tuned BERT performs quite well in detecting complex sensitive information. In [11] authors employ a logical-symbolic approach and introduce a frame-based knowledge graph specifically constructed to encompass personal data categories. This knowledge graph is developed by logically integrating pre-existing frames and has undergone evaluation as background knowledge within a Sensitive Information Detection system, utilizing a labeled dataset of sensitive information. The results are quite satisfactory. In [13] authors propose a deep learning approach for identifying private information in text and introduce a sequence labeling model based on the RoBERT neural network. Results demonstrate that the mode achieves quite satisfactory performance on Chinese text.

3. Empirical Study

A dataset was created to include user-generated textual data from pilot studies undertaken in the context of the EU-funded project WeNet. Users interacted and posed questions to a chatbot over a period of several days. All the exchanged messages were collected and archived, and an Ethics expert labeled each message based to indicate whether its content was deemed sensitive or not. The resulting labeled dataset consists of 1102 instances, 283 of which are labeled as sensitive, and 819 of which are labeled as non-sensitive. An additional 88 synthetic data instances were added to the dataset that belonged to the sensitive class label. The dataset was split 60%-20%-20% into a training set, a validation set, and a testing set, and various machine learning methods were trained and tested, using the validation set to fit their hyperparameters. Table 1 shows the testing performance of the resulting trained models for the best choice of hyperparameters of each machine learning method. The best performance was given by the SVM method, followed closely by the Naïve Bayes method.

Table 1. Testing performance of trained models using different machine learning methods.

Method	F1	Precision	Recall
Naïve Bayes	77.63	76.26	79.97
Decision Tree	64.67	60.18	67.91
Logistic Regression	73.19	72.47	75.78
SVM	77.66	77.44	78.92
k-Nearest Neighbor	50.26	48.06	69.61
Random Forest	69.32	68.46	72.18

We performed another experiment by adjusting the preprocessing in order to fix the data imbalance. We implement the SMOTE algorithm in order to oversample the minority class. A simple solution to imbalanced data is oversampling. The benefit of SMOTE is that it does not create duplicate data points, but rather creates synthetic data points that deviate slightly from the original data points. To rebalance the original training set, the SMOTE method implements an oversampling strategy. Instead of performing a simple replication of minority class instances, synthetic examples are the central concept of SMOTE. This new data is generated by interpolating across many occurrences of minority classes within a particular neighborhood. Because of this, the technique is said to be centered on the "feature space" rather than the "data space"; in other words, the algorithm is based on the values of the features and their relationships, as opposed to the data points as a whole. This has also led to an in-depth analysis of the theoretical relationship between original and synthetic instances, including the dimensionality of the data. Some features, like variance and correlation in the data and feature space, as well as the link between the distributions of training and test samples, are taken into account [3].

Table 2. SMOTE method

Method	Features	F1	Precision	Recall
SVM	Baseline	75.06	74.97	76.41
SVM	SMOTE on Minority Class	77.74	78.14	77.33

The results are quite encouraging and indicate that SMOTE algorithm can further enhance the performance of our methods improving the precision by almost 3%.

4. Explainability

Interpretability is important and desired in learning systems since it can provide valuable information about the way that predictions were made [4]. Existing attribute techniques are used to offer insight into what parts of questions affect sensitiveness. Also, we need our models to be free of bias and also fair, reliable, safe, and trustworthy, attributes that can be achieved and guaranteed by highly interpretable models.

LIME is a quite appropriate framework to use for local feature explainability. In a similar way, we also implemented SHAP and introduced its force plots into the comparison. Regarding SHAP in particular, in order to make the force plot more readable we only print the feature names for features the magnitude of their SHAP value is larger than min_perc * (sum of all abs shap values), where min_perc is declared as 0.08.

An example is illustrated in the textual data “How are you coping with your mental health?” which is annotated by the ethics expert to refer to sensitive data. Below, we present the local explanation for the sentence as well as the prediction probabilities. The determined probability to belong to the sensitive class was calculated to be 87.2%.

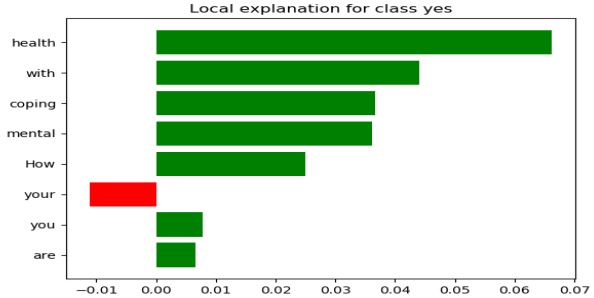

Figure 1. LIME features

As we can see from the above instance, our proposed model correctly recognizes the sentence as having the correct class - "Yes" in this case - with a strong confidence of 87.2%. Words such as "health" and "coping" heavily impact the prediction in favor of the "Yes" class with impact factors of 0.07 and 0.04 respectively. As expected, words referring to mental health should lead to a sensitive sentence prediction.

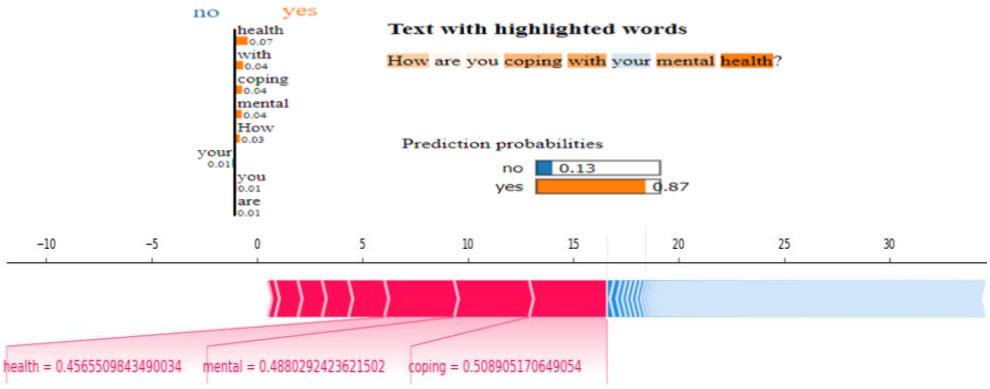

Figure 2. LIME and SHAP attribute graphs

In Figure 2, a more extensive overview of the LIME is illustrated, where we present with a weighted coloring the entire sentence. The variations of blue correspond to the non-sensitive data class, while variations of orange correspond to the sensitive data class. In the bottom graph, we observe the output of SHAP's force plot. It showcases words that are used in the particular instance, in an additive force layout from right to left. The word with the highest impact is the word *coping* with an impact of 0.50 on the decision made by the model.

5. Conclusions

Sensitive user content needs special handling in social networks. The recognition of sensitive content in users' interactions is quite important and can secure smoother user interactions, empower user inclusion and enhance the overall diversity awareness of the network. It can, also, assist in facing hate speech generation and discrimination issues. In this paper, we present a work on recognizing sensitive data in human interactions. Machine learning models were trained and tested on a real-life created dataset. The results indicate that the problem is feasible and can be automated. Our models are interpretable with the help of LIME and SHAP which offer insight on what aspects of user sentences affect sensitiveness and are consistent and robust predictors of sensitive content.

Acknowledgment

This work was supported by funding from the EU's Horizon 2020 Research and Innovation Programme under grant agreements no. 739578 and no. 823783, and from the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation, and Digital Policy. The authors are grateful to Laura Schelenz for discussions and for her contribution in tagging the data.

References

[1] Liberatore Frederico, & Quijano-Sanchez, Lara. What do we really need to compute the Tie Strength? An empirical study applied to Social Networks. Computer Communications, 110, 59, 2017

[2] Arnaboldi Valerio, Passarella, Andrea, Conti, Marco, & Dunbar, Rim, Online social networks: human cognitive constraints in Facebook and Twitter personal graphs. Elsevier. 2015
[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357
[4] Doshi-Velez, F. B. Kim, Towards a rigorous science of interpretable machine learning, 2017, arXiv preprint arXiv:1702.08608.
[5] Pappalardo, Luca, Rossetti, Giulio., & Pedreschi, Dino. " How Well Do We Know Each Other?" Detecting Tie Strength in Multidimensional Social Networks. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 1040-1045). IEEE, 2021
[6] Dunbar Rim., Arnaboldi, Valerio., Conti, Marco., & Passarella, Andrea. 2015. The structure of online social networks mirrors those in the offline world. Social networks, 43, 39-4
[7] Cunha, M., Mendes, R., & Vilela, J. P. (2021). A survey of privacy-preserving mechanisms for heterogeneous data types. Computer science review, 41, 100403.
[8] Sánchez, D., Batet, M., & Viejo, A. (2012). Detecting sensitive information from textual documents: an information-theoretic approach. In Modeling Decisions for Artificial Intelligence: 9th International Conference, MDAI 2012, Girona, Catalonia, Spain, November 21-23, 2012. pp. 173-184. Springer
[9] Xu, G., Qi, C., Yu, H., Xu, S., Zhao, C., & Yuan, J. (2019, October). Detecting sensitive information of unstructured text using convolutional neural network. In *2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)* (pp. 474-479). IEEE.
[10] Timmer, R. C., Liebowitz, D., Nepal, S., & Kanhere, S. S. (2021, December). Can pre-trained transformers be used in detecting complex sensitive sentences?-a monsanto case study. In *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications* pp. 90-97 IEEE.
[11] Gambarelli, G., & Gangemi, A. (2022). PRIVAFRAME: A Frame-Based Knowledge Graph for Sensitive Personal Data. *Big Data and Cognitive Computing*, *6*(3), 90.
[12] Majeed, A., Khan, S., & Hwang, S. O. (2022). A Comprehensive Analysis of Privacy-Preserving Solutions Developed for Online Social Networks. *Electronics*, *11*(13), 1931.
[13] Ning, Y., Wang, N., & Liu, A. (2021, April). Deep Learning based Privacy Information Identification approach for Unstructured Text. In *Journal of Physics: Conference Series* (Vol. 1848, No. 1, p. 012032). IOP Publishing.