HHAI2023

Sensitive Content Recognition in Social Interaction Messages

Isidoros Perikos and Loizos Michael

OPEN
UNIVERSITY OF
CYPRUS
www.ouc.ac.cy

Open University of Cyprus

Nicosia, Cyprus

Outline

- Introduction
- Methodology
- Evaluation Results
- Conclusions

Outline

- Introduction
- Methodology
- Evaluation Results
- Conclusions

Introduction

- The proliferation of social networks has increased our capacity to interact, communicate, and network, by creating new online environments to facilitate user interactions
- User interactions come mainly through the exchange of textual data, rich in personal information, opinions, and sentiments.

Introduction - Motivation

- 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 minorities and marginalized groups from being attacked by others.
- 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
- the detection of sensitive data can assist in facing hate speech and discrimination too.

Introduction - Contribution

- We designed and developed machine learning models to detect sensitive content
- We 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.

Outline

- Introduction
- Methodology
- Evaluation Results
- Conclusions

Methodology – Dataset creation

- 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.
- 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.

Methodology – ML Methods

- Various Machine Learning methods were designed and implemented
  - Naïve Bayes
  - Decision Tree
  - Logistic Regression
  - SVM
  - k-Nearest Neighbor
  - Random Forest

Methodology – SMOTE algorithm

- Imbalance data and problem since the minority of data in social networks contain sensitive information.
- SMOTE algorithm to face imbalance problem oversample the minority class.
- 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.

Methodology – Explainability

- Attribute techniques are used to offer insight into what parts of questions affect sensitiveness.
- 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 and SHAP are used as appropriate frameworks for local and global feature explainability.

Methodology – SMOTE algorithm

- 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

Outline

- Introduction
- Methodology
- Evaluation Results
- Conclusions

Evaluation – Training procedure

- 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.
- validation set, indicates how models are performing while training the model.

Evaluation – ML Results

- Results of the classifiers performance in detecting sensitive data

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

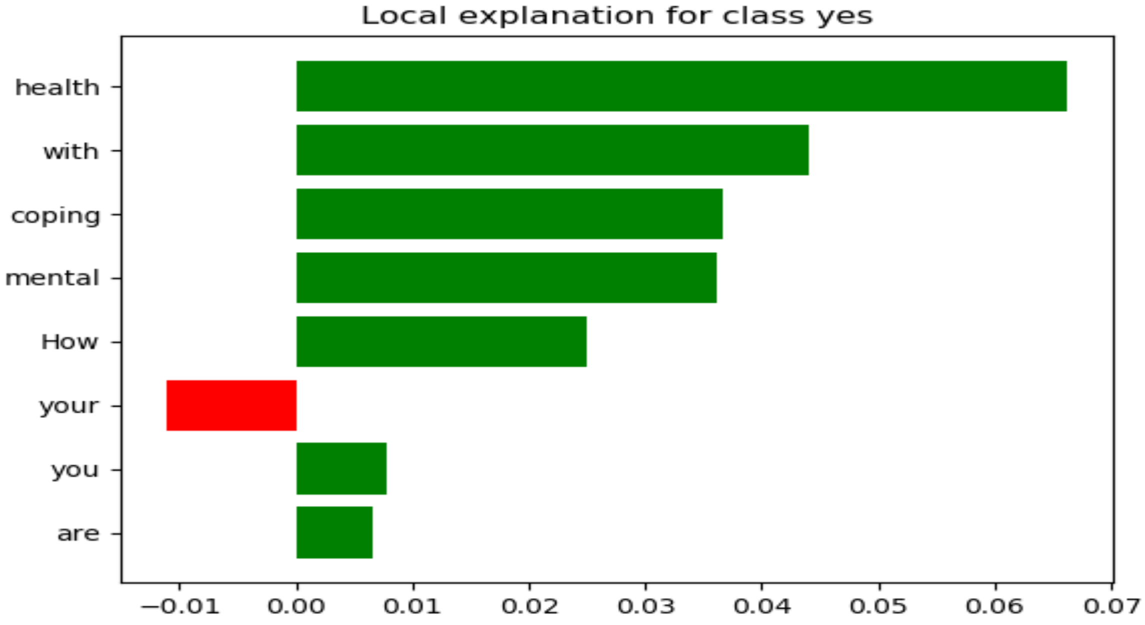
Evaluation – ML Results

- Results of the SMOTE integration on the top performing ML method that is the SVM

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

Evaluation – Interpretability Example case

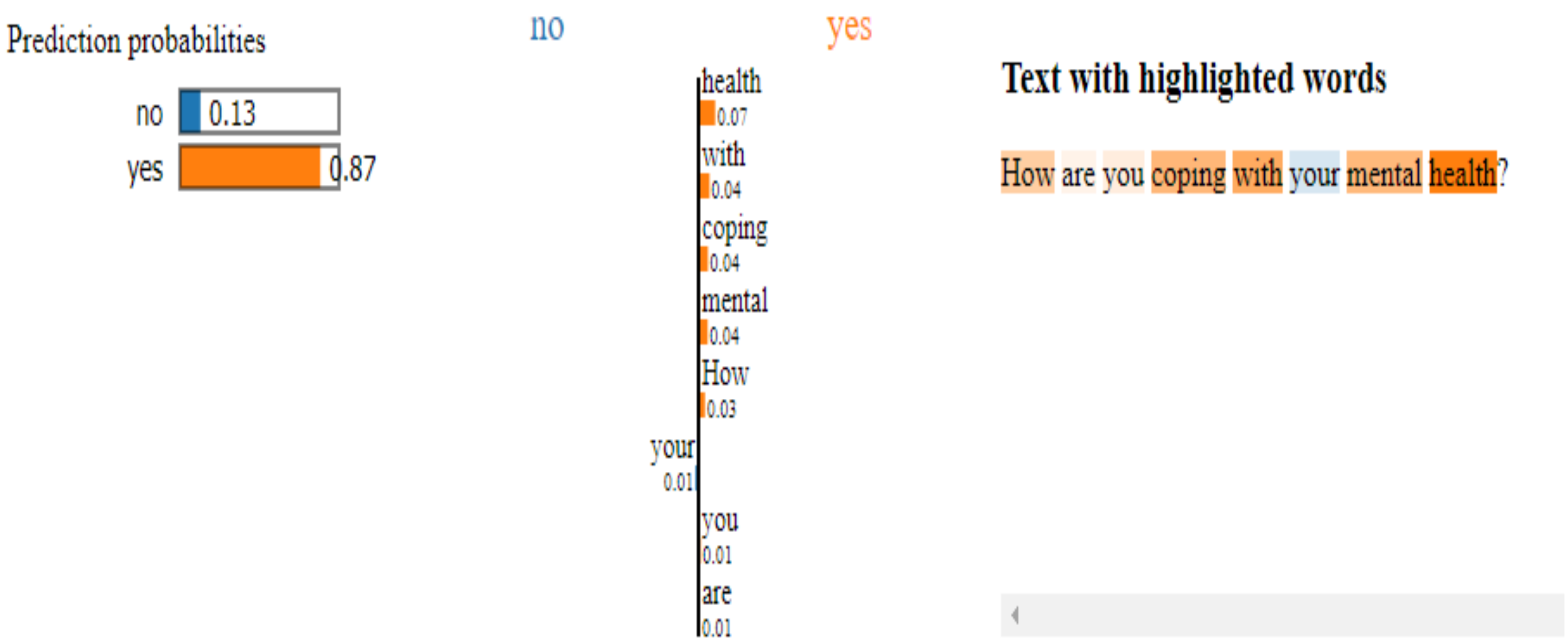
- “How are you coping with your mental health?” annotated to refer to sensitive personal data by ethics expert
- The determined probability to belong to the sensitive personal class was calculated to be 87.27



Evaluation – Interpretability Example case

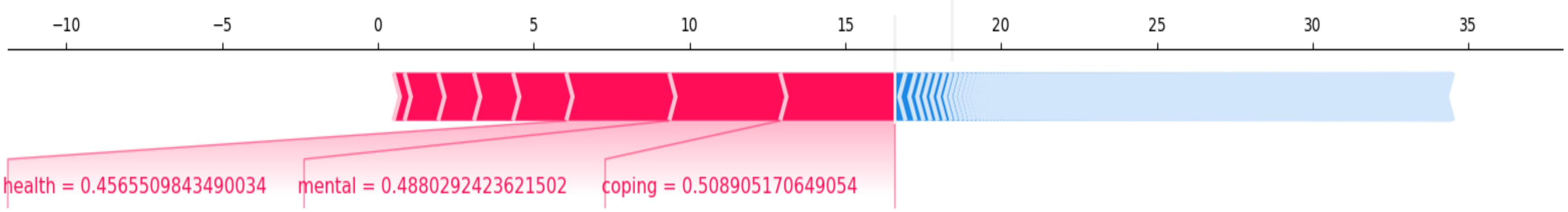
- The model correctly recognizes the sentence as having the correct class - “Yes” in this case - with a strong confidence of 87.27%.
- Words such as “health” and “coping” heavily impact the prediction in favour of the “Yes” class with impact factors of 0.07 and 0.04 respectively.
- As expected, words referring to mental health should lead to a sensitive sentence prediction.

Evaluation – Interpretability Example case



Text with highlighted words

How are you coping with your mental health?



Evaluation – Interpretability Example case

- LIME and SHAP-based experiments. In the first figure we find a more extensive overview of the LIME experiment, where we are presented with a weighted coloring for the entire sentences.
- The variations of blue correspond to the “No” class, while variations of orange correspond to the “Yes” class. On the other hand, in the second figure we can observe the output of SHAP’s force plot. It showcases all 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.

Outline

- Introduction
- Methodology
- Evaluation Results
- Conclusions

Conclusions

- Sensitive user content needs special handling in social networks.
  - secure smoother user interactions, empower user inclusion and enhance the overall diversity awareness of the network.
  - face hate speech and discrimination issues
- 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.
- Interpretability with the help of LIME and SHAP
  - insight on what aspects of user sentences affect sensitiveness and are consistent and robust predictors of sensitive content.

Thank you!