Journées de méthodologie statistique de l'Insee 2022

DETECTION DES INFRACTIONS RELEVANT DE LA CYBERDELINQUANCE A PARTIR D'UNE ANALYSE TEXTUELLE DES MANIERES D'OPERER

Maëlys BERNARD, Zoé GALLOS

Ministère de l'Intérieur, Service statistique ministériel de la sécurité intérieure

maelys.bernard@interieur.gouv.fr
zoe.gallos@interieur.gouv.fr

Mots-clés *(6 maximum)* : cybercriminalité, analyse textuelle, machine Learning, word embedding, Gradient Tree Boosting, réseau de neurones

Domaine concerné : *Machine learning*/Apprentissage statistique, classification, *NLP*/Analyse textuelle

---

Résumé (*entre 350 et 900 mots environ*)

La cybercriminalité est une nouvelle forme de criminalité en plein essor, et la lutte contre celle-ci est un enjeu important pour les services de police et de gendarmerie. Elle s'avère toutefois particulièrement difficile à quantifier. Elle a été définie par un groupe de travail piloté par le Service statistique ministériel de la sécurité intérieure (SSMSI) en 2014 comme regroupant toutes les infractions pénales tentées ou commises à l'encontre ou principalement au moyen d'un système d'information et de communication (SIC). La cybercriminalité n'est donc pas un champ infractionnel dont les contours seraient définis uniquement par des natures d'infractions. C'est surtout le mode opératoire qui fera d'une infraction qu'elle est qualifiée de « cyber » ou non. Dès lors, le phénomène est approché en utilisant des variables annexes associées aux infractions et caractérisant celles-ci : la coche « cyber » pour la gendarmerie, le mode opératoire, le contexte de la procédure, ainsi que la nature de lieu de l'infraction pour la police. Néanmoins, ces variables permettant de repérer la cybercriminalité ne sont pas toujours bien renseignées, et reposent largement sur l'interprétation du gendarme ou du policier. Ainsi, avec ces seules informations, le SSMSI n'est aujourd'hui pas en mesure de diffuser un indicateur fiable de la cybercriminalité.

Afin de pallier ce problème et de mesurer plus finement la cybercriminalité, un travail d'analyse textuelle appuyé par des techniques de *machine learning* a été implémenté en collaboration avec le SSP Lab. Cette analyse repose sur le texte contenant la description de la manière d'opérer, qui comporte quelques dizaines de mots. Ces manières d'opérer sont des résumés des « affaires » et sont remplies par les gendarmes et les policiers dans les logiciels

de rédaction des procédures. La saisie de cette zone textuelle étant obligatoire en gendarmerie, on se limite à ce champ dans un premier temps.

La labellisation d'un échantillon d'infractions sera menée par des experts-métier, allant au-delà de la simple distinction entre « cyber » et « non-cyber » et cherchant à distinguer une typologie propre à la cyberdélinquance, identifiant par exemple ce qui relève de moyens numériques pour des infractions de droit commun ou ce qui relève au contraire d'attaques purement cybercriminelles (rançongiciels, contenus illicites sur le web, etc .).

Les travaux préparatoires ont été menés en utilisant comme apprentissage une indicatrice croisant la coche « cyber » et des informations sur les natures d'infraction. Ils permettent de dérouler l'ensemble du processus de traitement qui comporte plusieurs étapes. Il est nécessaire dans un premier temps de rendre le texte utilisable pour l'analyse : le texte a donc été découpé en *token* par des étapes de prétraitements.
Le passage du texte en valeur numérique est réalisé par *word embedding* : l'algorithme *Word2Vec* est ici utilisé. La projection du document est ensuite réalisée par la moyenne des vecteurs projetés associés aux mots composant chaque document. L'apprentissage est ensuite réalisé sur le jeu de données test.
Enfin, une méthode de *machine learning* est appliquée pour prédire la cybercriminalité. Pour cela, plusieurs méthodes ont été testées telles que la régression logistique, les forêts aléatoires, les réseaux de neurones ou encore la méthode du *Gradient Tree Boosting*.
Toutes ces méthodes ont donné des résultats similaires quant à l'estimation de la cybercriminalité au sein des procédures enregistrées par la gendarmerie.
L'algorithme prédit à plus de 95% les infractions comme relevant de la cybercriminalité au sens de la définition. De même, il prédit moins de 1% d'infractions comme ne relevant pas de la cybercriminalité au sens de notre définition. La validation des résultats de l'algorithme sera par ailleurs complétée par une analyse d'un échantillon de procédures (prédites comme cyber et prédites comme non-cyber) par des personnels métier (policiers ou gendarmes).

Bibliographie

[1] Bird S., Klein E., Loper E., Natural Language Processing with Python, *O'Reilly*, Section 3.7 Regulars Expression for Tokenizing Text, pp 109-112 , juin 2009

[2] Biau G. Cadre B., Advances in Contemporary Statistics and Econometrics, *Editors Daouia A., Ruiz- Gazen A., Springer*, Chapter Optimization by Gradient Boosting pp.23-44, juin 2021.

[3] Mikolov T., Chen K., Corrado G., Dean J., Efficient Estimation of Word Representations in Vector Space, *arXiv:1301.3781v3 [cs.CL]*, 7 Sep 2013

[4] Razafindranovona T., Moreau A. Les défis de la mesure statistique de la cybercriminalité, *Revue de la Gendarmerie Nationale*, n°266, 4.trimestre 2019, janvier 2020

[5] Viano E. C. & al., Cybercrime, Organized Crime, and Societal Responses, *Editors Viano E. C.*, *Springer International Publishing*, Chapter Cybercrime: Definition, Typology, and Criminalization, pp2-33, 2017.

[6] Chollet F., Deep Learning with Python, *Manning Edition*, Part 1 Chapter 3 pp56-60 & Part 2 Chapter 6 pp178-195, 2017

JMS 2022 : Nomenclature des thématiques pour la classification des communications

Thématique	Sous-thématique
1. Théorie des sondages amont	1.1 Échantillonnage 1.2 Échantillonnages particuliers : spatial, équilibré, sur population continue .. 1.3 Bases de sondage 1.4 Unités statistiques
2. Théorie des sondages aval	2.1 Pondération et repondération, calage sur marges 2.2 Calcul de précision, estimation de variance
3. Contrôle et redressement des données, *data editing*	3.1 Non-réponse 3.2 Imputation 3.3 Identification, traitement des valeurs atypiques ou extrêmes (cas du milliardaire dans une enquête..) ou des valeurs influentes 3.4 Codification automatique
4. Collecte de données d'enquêtes	4.1 Protocole, conception des enquêtes, couverture de populations particulières 4.2 Conception de questionnaire 4.3 Multimode 4.4 Effets de mode 4.5 Effets d'oubli / de mémoire 4.6 Paradonnées
5. Combinaison de sources	5.1 Appariements, fusion de sources (*record linkage*), couplage « exact » de fichiers 5.2 Appariements probabilistes
6. Données administratives	6.1 Constitution de registres . 6.2 Signes de vie 6.3 Registres d'individus et de logements 6.4 Nettoyage des données administratives.
7. Statistique spatiale	7.1 Économétrie spatiale 7.2 Statistiques locales 7.3 Estimation sur petits domaines, carroyage 7.4 Zonages
8. Économétrie	8.1 Théorique 8.2 Appliquée (étude de cas) 8.3 Évaluation des politiques publiques
9. Modélisation	9.1 Mathématique ou stochastique 9.2 Microsimulations 9.3 Algorithmes
10. Séries temporelles	10.1 Analyse des séries 10.2 Désaisonnalisation 10.3 *Nowcasting* 10.4 Projections, prévisions
11. Science des données	11.1 Analyse des données, statistique descriptive, analyse factorielle 11.2 "*Big data*", nouvelles données, données massives

	11.3 *Machine learning* / Apprentissage statistique, classification 11.4 Intelligence artificielle, *deep learning*, réseaux de neurones 11.5 *NLP*, analyse textuelle 11.6 *Data viz*
12. Institutionnel, *open science*	12.1 Histoire 12.2 Confidentialité 12.3 Anonymisation 12.4 Diffusion 12.5 Communication 12.6 *Open data*, documentation, métadonnées, mise à disposition, partage de codes, reproductibilité
13. Comptabilité nationale	13.1 Nationale 13.2 Estimations régionales 13.3 Indicateurs macro-économiques
14. Concepts et mesures	14.1 Indicateurs, échelles, indices 14.2 Incertitudes 14.3 Erreurs de mesures
15. Mesures et impact de la pandémie de Covid	
16. Données médicales	
17. Démographie	17.1 Recensement 17.2 Indicateurs ou études démographiques
18. Statistique d'entreprises	
19. Enseignement, éducation	19.1 Évaluation des élèves 19.2 Autre
20. Autre	