**CYBER INCIDENT REPORTS: E X T R A P O L AT I N G S E V E R I T Y USING NEURAL NETWORKS**

**Note de synthèse**

Cyber incident reports: Extrapolating severity unsing neural networks

**Problème** : disponibilité des données fiable sur les attack cybers pour modéliser le phénomène, en particulier tarifer et évaluer les réclamations liées au contrat d’assurance risque cyber.

**Objectif :** Face à la pénurie de données fiables sur le risque cyber, cette étude propose une méthode pour estimer la sévérité d’un incident cyber à partir de la description textuelle, en utilisant des réseaux de neurones et des techniques de traitement du langage naturel (NLP).

Le document montre l’importance d’avoir une source de données assez vaste pour pouvoir modéliser la sévérité du risque cyber. Ici il explique comment extraire les informations des sources textuelles (Benchmark database) en utilisant le NLP (Neural Langage Processing) pour prédire le montant ou au moins un indicateur de la sévérité. En fin le modèle est utilisé pour estimer la sévérité des attaques cybers dont on a seulement la description.

**Section 1 : Présentation des sources de données fiable pour extraire données d’attaques cyber utilisées pour calibrer le modèle**

**Description de la base PRC (Privacy Rights Clearinghouse)** :

La PRC est une association au Etats Unis dont le but est d’éduquer la population sur les problèmes liés à leur vie privée. La raison de la popularité de cette base est qu’elle contient quelques marqueurs de la sévérité (number of records, number of accounts that are exposed.

Le lien entre « number of records » et la vraie perte économique à été démontré par Ponemon (2018), Farkas (2021). Toutefois, compte tenu de l’irrégularité des modèles de lien, on se concentre dans ce rapport uniquement sur le « number of records » comme proxy de la sévérité des attaques cybers qui sera utilisé dans la suite pour la modélisation la sévérité.

Une autre limite de cette base de données est qu’elle est uniquement concentrée sur les Etats Unis. (Prendre en compte le fait que le nombre d’attaques cybers venant de l’extérieur se produisant au Etats Unis est relativement faible.

La Base de données ne se refaire qu’au risque « data breach »

**Statistique descriptive**

**Analyse de variables**

24% des données de la base contiennent une variable « number of records » égale 0 ce qui seront supprimé, réduisant ainsi la taille de la base.

Les données sont subdivisées comme suit : 40 % des pertes les plus importantes soit 40% des observation ayant le « number of records ».

Vu que la variable « number of record » est notre variable d’intérêt nous a étudions sa distribution par quartiles, pour déterminer le seuil à partir duquel les pertes sont importantes.

(C’est quoi la notion de perte importante ? )

* Analyse de la distribution des pertes
* Analyse du nombre de réclamation par année
* Analyse du nombre de réclamation par type d’organisation (domaine de médecine et le plus touché). (Ceci dit, nous nous attendons à ce que le champ lexical de la médecine domine dans les rapports de réclamation d’attaque cyber PRC).
* Nombre de réclamation par type de brèche
* Analyse des réclamations selon les mots dans les textes

**Section 2 : Text embedding et Réseaux de Neurones**

Description du perceptron multicouche classique :

**Cadre des réseaux de neurones**

C’est un modèle de machine learning caractérisé par un grand nombre de paramètre (faible approximation des erreurs d’estimation) et un algorithme efficient pour estimer les valeurs de ces paramètres. Ceci rend ce modèle plus compétitif en assurance.

Une image contenant ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

**Word embedding**

Procédure utilisée pour convertir du texte en vecteur . Premièrement convertir un mot en nombre : one-hot encoding, suivie d’une projection dans un espace de petite dimension où deux mots de signification proche devraient réprésentés par deux vecteurs qui sont proche (Word2vec : methode utilisant les réseaux de neurone artificiel pour transformer les mots en vecteur, Word2vec is a type of shallow two-layer neural network designed to generate word embeddings, which are vector representations of words that capture contextual information).

Comme extention nous avons des algorithmes d’embedding word :

CNN (Convolutional Neural Networks)

LSTM (Long-Short Term Memory)

**Section 3) Application sur le PRC data**

Dans cette section nous définissons un seuil égal au quantile de la variable « Number of record » tel que à 40% des observation au-dessus de ce seuil on le plus grand « Number of record ». Ce seuil permet de séparer notre échantillon de données en deux blocs, Nous considérons un problème de classification binaire, affectant 1 aux incidents ayant un « Number of record » supérieur au seuil défini, et 0 au reste.

**L’architecture et les hyperparamètres du réseau de neurone**

**Approche et outils**

Premièrement les données textuelles sont nettoyées en retirant les caractères relatifs aux dates, ponctuation, et les mots (stop-word) qui ne fournissent aucune information en utilisant la librairie NTLK. En suite procède à la tokenisation avant de créer les bigrams et les trigrams avec la librairie GENSIM, En fin nous Utilisons le model Gensim’s Word2vec pour transformer les mots en vecteur ayant une certaine similarité. En fin nous utilisant la Librairie KERAS pour créer le Perceptron.

**Challenges rencontrés durant l’analyse**

La répétition des descriptions de certaines données textuelles qui renvoient les mêmes probabilités, ce qui n’améliore pas les performances du modèle, par conséquent nous les avons supprimés pour remédier au problème.

**Qualité de la prédiction**

Dans un premier temps nous divisons notre base en échantillon d’entrainement et de test, les résultats sont présentés par la suite.

(Y’a-t-il un échantillon de validation ?)

**Comment les paramètres ont été fixés**

On effectue une validation croisée pour déterminer le nombre de couche intermédiaires et de neurones dans chaque couche. (Comment ?)

Ici le modèle retenu est celui qui maximise le F1 score qui est la moyenne harmonique entre la précision et le rappel

Une image contenant texte, Police, capture d’écran, document

Le contenu généré par l’IA peut être incorrect.

**Résultat**

Après validation croisée on a le résultat optimal suivant :

Une image contenant diagramme, ligne

Le contenu généré par l’IA peut être incorrect.

En utilisant l’échantillon de test, la matrice de confusion avec le F1 score est données par :

Une image contenant texte, Police, capture d’écran, nombre

Le contenu généré par l’IA peut être incorrect.

Pour évaluer le niveau de prédictibilité de notre modèle nous analysons la figure suivante :

Une image contenant texte, capture d’écran, ligne, Police

Le contenu généré par l’IA peut être incorrect.

Une image contenant capture d’écran, texte, diagramme, carré

Le contenu généré par l’IA peut être incorrect.

Pour les 1102 observations de l’échantillon de test, les graphiques ci-dessous présentent les taux de sinistres graves dans l’échantillon de test ainsi que la précision selon les modalités. Les figures 12 et 13 correspondent aux prédictions de sinistres graves, tandis que les figures 14 et 15 concernent les prédictions de sinistres peu graves (attritionnels).

Une image contenant texte, capture d’écran, Police, diagramme

Le contenu généré par l’IA peut être incorrect.

**Types d’organisations bien prédits** :

* EDU (Établissements éducatifs)
* BSO (Autres entreprises)
* BSF (Entreprises financières et assurances)
* BSR (Commerces et e-commerce)
* **Plus de 65 %** des pertes graves sont bien identifiées.

**Cas médical (MED) mal prédit** :

* Les sinistres dans le secteur médical sont **moins bien détectés**.
* Vocabulaire trop **spécifique et répétitif**, donc **peu révélateur** de la gravité.
* L’**ampleur** du sinistre est sans doute plus liée à la **taille de la fuite** qu’aux mots utilisés.

**Types de fuites bien prédits** :

* **Piratage (hack)**
* **Périphériques portables perdus ou volés** (clés USB, ordinateurs portables, CD, disques durs, etc.)

**Utilisation des expressions régulières (regex)**

L’objectif ici est d’améliorer la précision des prédictions du modèle de réseau de neurones en ajoutant une étape complémentaire avec des expressions régulières. En effet, un réseau de neurones peut parfois rater des cas graves, créant ainsi des faux négatifs (le modèle dit "pas grave", alors que c'est grave). Les regex vont aider à détecter ces cas manqués.

**Comment ça fonctionne ?**

1. On a un modèle (ici un **perceptron multicouche**) qui prédit si une description d’incident est grave ou non.
2. Ensuite, **on applique une analyse avec des regex** pour **repérer certains mots-clés, motifs ou formats spécifiques** (exemples : emails, URL, termes comme "hacked", "ransomware", "stolen", etc.).
3. Si un cas a été **mal classé par le modèle**, mais que la regex détecte un **motif révélateur de gravité**, on peut **corriger la prédiction**.

**re** est la librairie standard en Python pour les expressions régulières.

**Analyse**

**Méthode :**

* On extrait des expressions *nombre + mot* avec des regex.
* On garde seulement celles dont le mot est dans une liste importante (V).
* On somme les nombres associés à ces mots.
* Si le total dépasse un seuil, on considère que la réclamation est grave.
* Cela permet d’améliorer la détection automatique des sinistres importants, même si le réseau de neurones ne les avait pas détectés.

(Comment définir le seuil dans ce cas ?) à dire d’expert ?

Etape préliminaire (nettoyage des données)

Choix du vocabulaire (qui définir les sinistres sévères) : V= {Patient, people, student, employee, customer}. Ces mots désignent des personnes réelles, donc le nombre devant eux reflète souvent directement l’ampleur de l’impact (ex : "27000 employees" = grave).

**Comment définir le seuil pour le Regex ?**

On fait varier le seuil **(threshold)** pour décider à partir de quelle somme, on classe un sinistre comme **grave**, on mesure la performance de la classification (via une matrice de confusion). On cherche à maximiser le F1-score (un bon équilibre entre précision et rappel).

Une image contenant texte, ligne, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Le seuil optimal trouvé dans notre cas est 4700 (la somme des nombres doit dépasser 4700 pour considérer un sinistre grave).

**Résultats**

Une image contenant texte, capture d’écran, Police, diagramme

Le contenu généré par l’IA peut être incorrect.

On peut dire que :

* Les regex permettent de corriger certaines erreurs du modèle.
* Le F1 score passe à 66%, ce qui est une nette amélioration.
* La clé est de choisir un bon seuil et un vocabulaire pertinent pour interpréter les chiffres.
* Mais la méthode ne couvre pas tous les cas, car certaines descriptions sont trop vagues ou non numériques.

Une image contenant texte, capture d’écran, Police, ligne

Le contenu généré par l’IA peut être incorrect.

**Analyse des données manquantes du PRC**

Utilisation de l’algorithme présenté supra pour prédire le « number of record » qui est maquant.

Statistique de comparaison avec la base de données complète :

Une image contenant texte, capture d’écran, Police, Tracé

Le contenu généré par l’IA peut être incorrect.

* Les institutions médicales (MED) et gouvernementales (GOV) doivent être transparentes et signaler leurs incidents, donc elles ont moins de données manquantes.
* Les entreprises business (BSO, BSF, BSR) sont souvent moins transparentes ou moins précises dans leurs diagnostics, donc elles ont plus de données manquantes (peut-être parce qu’elles ne savent pas combien de dossiers ont été affectés ou ne veulent pas le révéler).
* Les violations comme les hacks et les violations physiques sont plus fréquentes dans la base de données incomplète.

**Proportion of severe claim rate of each modality for the chosen mode**

**Une image contenant texte, capture d’écran, Police, Tracé

Le contenu généré par l’IA peut être incorrect.**

* Le taux de réclamations graves est plus bas pour MED, GOV et BSF dans la base de données incomplète que dans la base complète.
* À l’inverse, les institutions éducatives (EDU) et certains types d’entreprises (BSO, BSR) ont un taux de réclamations graves plus élevé dans la base incomplète.
* Les violations physiques (notées PORT et STAT) sont moins graves dans la base incomplète que dans la base complète.

**Conclusion**

Dans cette note, les auteurs ont présenté une méthodologie générique pour prédire la gravité d’un incident cyber à partir de son texte descriptif.

L’exemple utilisé porte sur la base de données de fuites de données (data breaches) PRC au USA. Cependant, la méthode peut s’étendre à d’autres types d’incidents cyber plus complexes et des bases de données plus riches.

En somme, cette méthode, simple mais efficace, montre que l’analyse du texte descriptif d’un incident cyber peut grandement aider à évaluer la gravité de cet incident, et ainsi faciliter la gestion et la prévision des risques dans le domaine de la cybersécurité.

**Question ?**

Une approche pour inclure dans nos modèles les variables telles que : secteur d’activité, taille de l’entreprise, budget cybersécurité

Comment peut-on faire pour prédire avec cette méthode des variables quantitatives (par exemple, le montant des pertes) ?