Exploiter les séries texto-temporelles en actuariat

Jean-Thomas Baillargeon, FICA, FSA, CERA, M.Sc.

Séminaire départemental de l'école d'actuariat de l'Université Laval

jean-thomas.baillargeon@ift.ulaval.ca

Avril 2023




Introduction

On s'intéresse au département des réclamations en assurance auto

Introduction

On s'intéresse au département des réclamations en assurance auto

- Front entre l'assureur et ses assurés

Introduction

On s'intéresse au département des réclamations en assurance auto

- Front entre l'assureur et ses assurés
- Satisfaction du règlement vs minimiser les coûts

Introduction

On s'intéresse au département des réclamations en assurance auto

- Front entre l'assureur et ses assurés
- Satisfaction du règlement vs minimiser les coûts
- Il n'existe pas toujours une seule solution

Dommages physiques, ce n'est pas trop dangereux

Dommages physiques, ce n'est pas trop dangereux

- Déterminer le coût de remplacement

Introduction

Dommages physiques, ce n'est pas trop dangereux

- Déterminer le coût de remplacement
- Négociations possibles
  - Risque de fuite (*leakage*)
  - Risque de réputation

Introduction

Dommages corporels, risques long terme à modéliser

Introduction

Dommages corporels, risques long terme à modéliser

- Coûts difficiles à évaluer
  - Non connu au moment du sinistre
  - Évolue dans le temps selon la réussite des traitements

Introduction

Dommages corporels, risques long terme à modéliser

- Coûts difficiles à évaluer
  - Non connu au moment du sinistre
  - Évolue dans le temps selon la réussite des traitements
- Coûts reliés
  - Traitements médicaux
  - Soins à domicile
  - Remplacement de revenus

Introduction

Dommages corporels, risques long terme à modéliser

- Coûts difficiles à évaluer
  - Non connu au moment du sinistre
  - Évolue dans le temps selon la réussite des traitements
- Coûts reliés
  - Traitements médicaux
  - Soins à domicile
  - Remplacement de revenus
- Potentiellement coûteux $>$1M \$

Introduction

Frein à la compréhension de l'évolution long terme des sinistres

- Données utilisées pour modéliser les dommages corporels
  - Données de tarification
  - Contexte incomplet de la réclamation

Frein à la compréhension de l'évolution long terme des sinistres

- Données utilisées pour modéliser les dommages corporels
  - Données de tarification
  - Contexte incomplet de la réclamation
- Données non structurées contiennent plus d'information - les notes de sinistres
  - Description de l'accident
  - Correspondances entre les intervenants durant le développement du sinistre

Une expertise importante joue un rôle central

- Nécessite une grande connaissance en ingénierie des données et logicielle.

Une expertise importante joue un rôle central

- Nécessite une grande connaissance en ingénierie des données et logicielle.
- Nécessite d'avoir des connaissances pratiques et théoriques en actuariat.

Introduction

Une expertise importante joue un rôle central

- Nécessite une grande connaissance en ingénierie des données et logicielle.
- Nécessite d'avoir des connaissances pratiques et théoriques en actuariat.
- Double expertise rare, difficile et coûteuse à développer.

Agenda

Exploiter les séries texto-temporelle en actuariat

- Étude de cas #1 : Estimer le nombre de blessés dans une description d'accident

Agenda

Exploiter les séries texto-temporelle en actuariat

- Étude de cas #1 : Estimer le nombre de blessés dans une description d'accident
  - Comparer la création manuelle d'attributs et la modélisation automatique avec l'apprentissage de représentations.
  - Expliquer les modèles grâce aux patrons textuels utilisés.

Agenda

Exploiter les séries texto-temporelle en actuariat

- Étude de cas #1 : Estimer le nombre de blessés dans une description d'accident
  - Comparer la création manuelle d'attributs et la modélisation automatique avec l'apprentissage de représentations.
  - Expliquer les modèles grâce aux patrons textuels utilisés.
- Étude de cas #2 : Détection des sinistres catastrophiques à l'aide d'un dossier de réclamation

Agenda

Exploiter les séries texto-temporelle en actuariat

- Étude de cas #1 : Estimer le nombre de blessés dans une description d'accident
  - Comparer la création manuelle d'attributs et la modélisation automatique avec l'apprentissage de représentations.
  - Expliquer les modèles grâce aux patrons textuels utilisés.
- Étude de cas #2 : Détection des sinistres catastrophiques à l'aide d'un dossier de réclamation
  - Modéliser de longues séquences texto-temporelles.
  - Comparer les explications de modèles.

Étude de cas : Estimer le nombre de blessés dans une description d'accident

Estimer le nombre de blessés dans un rapport d'accident

Mise en scène

- Les autorités gouvernementales commencent à diffuser en temps réel les descriptions d'accidents automobiles sur un portail de données ouvertes.

Estimer le nombre de blessés dans un rapport d'accident

Mise en scène

- Les autorités gouvernementales commencent à diffuser en temps réel les descriptions d'accidents automobiles sur un portail de données ouvertes.
- Vous aimeriez utiliser ces rapports afin de déterminer le nombre de personnes blessées et ainsi améliorer les réserves pour sinistres encourus, mais non reportés (IBNR) pour les réclamations en dommage corporel.

Estimer le nombre de blessés dans un rapport d'accident

Mise en scène

- Les autorités gouvernementales commencent à diffuser en temps réel les descriptions d'accidents automobiles sur un portail de données ouvertes.
- Vous aimeriez utiliser ces rapports afin de déterminer le nombre de personnes blessées et ainsi améliorer les réserves pour sinistres encourus, mais non reportés (IBNR) pour les réclamations en dommage corporel.
- **Problème** : Pas de données = pas de modèle.

Estimer le nombre de blessés dans un rapport d'accident

Mise en scène

- Les autorités gouvernementales commencent à diffuser en temps réel les descriptions d'accidents automobiles sur un portail de données ouvertes.
- Vous aimeriez utiliser ces rapports afin de déterminer le nombre de personnes blessées et ainsi améliorer les réserves pour sinistres encourus, mais non reportés (IBNR) pour les réclamations en dommage corporel.
- **Problème** : Pas de données = pas de modèle.
- **Solution** : Créer un modèle avec un jeu de données similaire comme base et le bonifier avec les données qui arrivent en temps réel.
  - National Highway Trafic Safety Association (NHTSA)
  - Présenté par [Borba, 2013]
  - Contient 6949 descriptions d'accident et le nombre de personnes blessées

Estimer le nombre de blessés dans un rapport d'accident

Ce qu'on tente de faire

Travaux reliés

- [Zappa et al., 2019] identifie manuellement les substances intoxicantes dans les descriptions d'accidents afin d'évaluer leur impact sur la sévérité de l'accident

Estimer le nombre de blessés dans un rapport d'accident

Ce qu'on tente de faire

Travaux reliés

- [Zappa et al., 2019] identifie manuellement les substances intoxicantes dans les descriptions d'accidents afin d'évaluer leur impact sur la sévérité de l'accident
- [Tixier et al., 2016] et [Baker et al., 2019] identifient manuellement des facteurs de risque dans des descriptions d'environnement de travail afin de modéliser la sévérité des accidents potentiels pouvant s'y produire.

Estimer le nombre de blessés dans un rapport d'accident

Ce qu'on tente de faire

Travaux reliés

- [Zappa et al., 2019] identifie manuellement les substances intoxicantes dans les descriptions d'accidents afin d'évaluer leur impact sur la sévérité de l'accident
- [Tixier et al., 2016] et [Baker et al., 2019] identifient manuellement des facteurs de risque dans des descriptions d'environnement de travail afin de modéliser la sévérité des accidents potentiels pouvant s'y produire.
- [Manski et al., 2021] utilise de brèves descriptions d'évènements météorologiques pour prédire les coûts du sinistre avec un nombre restreint de mots choisis automatiquement.

Estimer le nombre de blessés dans un rapport d'accident

Ce qu'on tente de faire

Défi

- Représenter automatiquement et efficacement les données textuelles.

Estimer le nombre de blessés dans un rapport d'accident

Ce qu'on tente de faire

Défi

- Représenter automatiquement et efficacement les données textuelles.
- Estimer le nombre de blessés dans un accident selon sa description.

Estimer le nombre de blessés dans un rapport d'accident

Ce qu'on tente de faire

Défi

- Représenter automatiquement et efficacement les données textuelles.
- Estimer le nombre de blessés dans un accident selon sa description.
- Similaire à [Baillargeon et al., 2020] et [Blier-Wong et al., 2021].

Estimer le nombre de blessés dans un rapport d'accident

Ce qu'on tente de faire

Comment transforme-t-on ceci :

> This two-vehicle crash occurred just before noon on an eight-lane, asphalt, straight, divided roadway.
>
> Vehicle one (V1), a 2004 Chevrolet Trailblazer SUV driven by a 51 year-old male with one passenger, was traveling west in lane two. Vehicle two (V2), a 1994 RTD Gillig Bus driven by a 50 year-old female with two passengers, was stopped in lane two in front of V1.
>
> The front of V1 contacted the back of V2. V2 came to rest at impact. V1 was towed due to frontal damage. V2 was driven from the scene.

En cela : 1

Estimer le nombre de blessés dans un rapport d'accident

Problème de régression

Régression d'un processus de comptage

- $\mathbf{X}$, représentation vectorielle des n documents d_i, où $i \in 1, ..., n$ du corpus $\mathcal{C}$

Estimer le nombre de blessés dans un rapport d'accident

Problème de régression

Régression d'un processus de comptage

- $\mathbf{X}$, représentation vectorielle des n documents d_i, où $i \in 1, ..., n$ du corpus $\mathcal{C}$
- $\mathbf{Y}$, la variable réponse (le nombre de blessés dans le i^e accident)

Comment créer $\mathbf{X}$?

- Les actuaires sont experts pour trouver des patrons dans des données.

Estimer le nombre de blessés dans un rapport d'accident

Variables explicatives

Comment créer $\mathbf{X}$?

- Les actuaires sont experts pour trouver des patrons dans des données.
- Les données textuelles sont variées et complexes :
  - Variabilité des orthographes (Typo, acronymes, abréviations).
  - Variabilité des styles d'écritures.
  - Dépendances entre les groupes syntaxiques complexes à modéliser.

Estimer le nombre de blessés dans un rapport d'accident

Variables explicatives

Comment créer $\mathbf{X}$?

- Les actuaires sont experts pour trouver des patrons dans des données.
- Les données textuelles sont variées et complexes :
  - Variabilité des orthographes (Typo, acronymes, abréviations).
  - Variabilité des styles d'écritures.
  - Dépendances entre les groupes syntaxiques complexes à modéliser.
- Les patrons présents ne peuvent pas être tous capturés manuellement.

Indices pour compter les blessés

This two-vehicle crash occurred just before noon on an eight-lane, asphalt, straight, divided roadway.

Vehicle one (V1), a 2004 Chevrolet Trailblazer SUV driven by a 51 year-old male with one passenger, was traveling west in lane two. Vehicle two (V2), a 1994 RTD Gillig Bus driven by a 50 year-old female with two passengers, was stopped in lane two in front of V1.

The front of V1 contacted the back of V2. V2 came to rest at impact. V1 was towed due to frontal damage. V2 was driven from the scene.

Indices pour compter les blessés

This two-vehicle crash occurred just before noon on an eight-lane, asphalt, straight, divided roadway.

Vehicle one (V1), a 2004 Chevrolet Trailblazer SUV driven by a 51 year-old male with one passenger, was traveling west in lane two. Vehicle two (V2), a 1994 RTD Gillig Bus driven by a 50 year-old female with two passengers, was stopped in lane two in front of V1.

The front of V1 contacted the back of V2. V2 came to rest at impact. V1 was towed due to frontal damage. V2 was driven from the scene.

Indices pour compter les blessés

This two-vehicle crash occurred just before noon on an eight-lane, asphalt, straight, divided roadway.

Vehicle one (V1), a 2004 Chevrolet Trailblazer SUV driven by a 51 year-old male with one passenger, was traveling west in lane two. Vehicle two (V2), a 1994 RTD Gillig Bus driven by a 50 year-old female with two passengers, was stopped in lane two in front of V1.

The front of V1 contacted the back of V2. V2 came to rest at impact. V1 was towed due to frontal damage. V2 was driven from the scene.

Estimer le nombre de blessés dans un rapport d'accident

Analyse manuelle d'un texte

Indices pour compter les blessés

This two-vehicle crash occurred just before noon on an eight-lane, asphalt, straight, divided roadway.

Vehicle one (V1), a 2004 Chevrolet Trailblazer SUV driven by a 51 year-old male with one passenger, was traveling west in lane two. Vehicle two (V2), a 1994 RTD Gillig Bus driven by a 50 year-old female with two passengers, was stopped in lane two in front of V1.

The front of V1 contacted the back of V2. V2 came to rest at impact. V1 was towed due to frontal damage. V2 was driven from the scene.

Limites des attributs générés manuellement

Limites des attributs générés manuellement

- Difficilement réutilisables pour d'autres tâches

Limites des attributs générés manuellement

- Difficilement réutilisables pour d'autres tâches
- Fragiles aux patrons n'ayant été jamais vus

Estimer le nombre de blessés dans un rapport d'accident

Analyse manuelle d'un texte

Limites des attributs générés manuellement

- Difficilement réutilisables pour d'autres tâches
- Fragiles aux patrons n'ayant été jamais vus
- Utilisent un sous-ensemble de l'information disponible

Estimer le nombre de blessés dans un rapport d'accident

Analyse manuelle d'un texte

Limites des attributs générés manuellement

- Difficilement réutilisables pour d'autres tâches
- Fragiles aux patrons n'ayant été jamais vus
- Utilisent un sous-ensemble de l'information disponible
- Non garant d'un modèle interprétable

Estimer le nombre de blessés dans un rapport d'accident

Approche par apprentissage de représentations

Apprendre automatiquement les attributs à utiliser avec un réseau de neurones ([Blier-Wong et al., 2021])

Estimer le nombre de blessés dans un rapport d'accident

Approche par apprentissage de représentations

Apprendre automatiquement les attributs à utiliser avec un réseau de neurones ([Blier-Wong et al., 2021])

- Réseau de neurones : Encodeur / Décodeur

Estimer le nombre de blessés dans un rapport d'accident

Approche par apprentissage de représentations

Apprendre automatiquement les attributs à utiliser avec un réseau de neurones ([Blier-Wong et al., 2021])

- Réseau de neurones : Encodeur / Décodeur
- Encodeur : génère un encodage (vecteur dense à dimensionnalité réduite) à partir des données brutes.

Estimer le nombre de blessés dans un rapport d'accident

Approche par apprentissage de représentations

Apprendre automatiquement les attributs à utiliser avec un réseau de neurones ([Blier-Wong et al., 2021])

- Réseau de neurones : Encodeur / Décodeur
- Encodeur : génère un encodage (vecteur dense à dimensionnalité réduite) à partir des données brutes.
- Décodeur : Réalise une tâche de prédiction avec l'encodage (et ajuste les poids du modèle)

Estimer le nombre de blessés dans un rapport d'accident

Approche par apprentissage de représentations

Apprendre automatiquement les attributs à utiliser avec un réseau de neurones ([Blier-Wong et al., 2021])

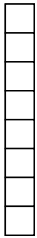
- Réseau de neurones : Encodeur / Décodeur
- Encodeur : génère un encodage (vecteur dense à dimensionnalité réduite) à partir des données brutes.
- Décodeur : Réalise une tâche de prédiction avec l'encodage (et ajuste les poids du modèle)
- Selon le paradigme d'attributs fais à la main
  - Encodeur = Actuaire
  - Décodeur = GLM

Estimer le nombre de blessés dans un rapport d'accident

Analogie importante : analyse par composante principale

Estimer le nombre de blessés dans un rapport d'accident

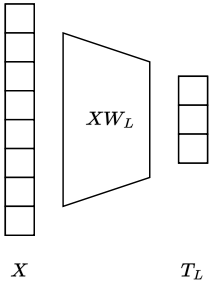
Analogie importante : analyse par composante principale



x

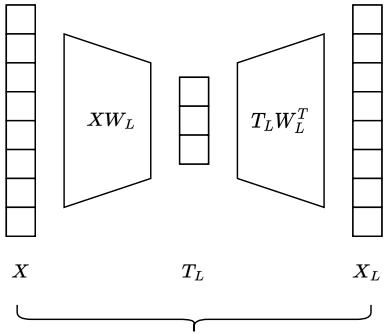
Estimer le nombre de blessés dans un rapport d'accident

Analogie importante : analyse par composante principale



Estimer le nombre de blessés dans un rapport d'accident

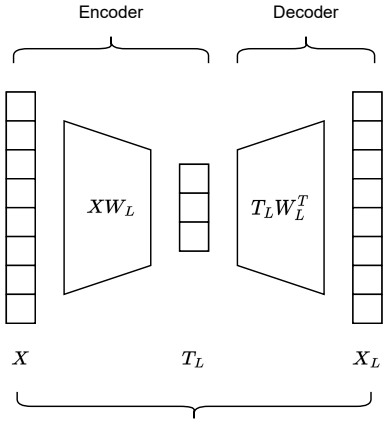
Analogie importante : analyse par composante principale



$$W_L = \text{argmin}(||X - X_L||_2^2)$$

Estimer le nombre de blessés dans un rapport d'accident

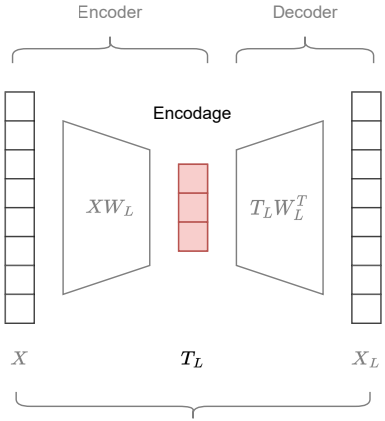
Analogie importante : analyse par composante principale



$$W_L = \text{argmin}(||X - X_L||_2^2)$$

Estimer le nombre de blessés dans un rapport d'accident

Analogie importante : analyse par composante principale



$$W_L = \text{argmin}(||X - X_L||_2^2)$$

Estimer le nombre de blessés dans un rapport d'accident

Modéliser les données brutes

Technique traditionnelle : représentation en sac de mots (Bag of Words, BoW)

- Document d_i : vecteur avec une position pour chaque mot distinct du corpus

Estimer le nombre de blessés dans un rapport d'accident

Modéliser les données brutes

Technique traditionnelle : représentation en sac de mots (Bag of Words, BoW)

- Document d_i : vecteur avec une position pour chaque mot distinct du corpus
- Clairsemé : rempli de 0, sauf 1 aux positions associées aux mots présents dans document d_i.

Estimer le nombre de blessés dans un rapport d'accident

Modéliser les données brutes

Technique traditionnelle : représentation en sac de mots (Bag of Words, BoW)

- Document d_i : vecteur avec une position pour chaque mot distinct du corpus
- Clairsemé : rempli de 0, sauf 1 aux positions associées aux mots présents dans document d_i.

On vectorise la phrase suivante :

```
V1 was towed from the scene and V2 was driven away.
```

Estimer le nombre de blessés dans un rapport d'accident

Modéliser les données brutes

Technique traditionnelle : représentation en sac de mots (Bag of Words, BoW)

- Document d_i : vecteur avec une position pour chaque mot distinct du corpus
- Clairsemé : rempli de 0, sauf 1 aux positions associées aux mots présents dans document d_i.

On vectorise la phrase suivante :

```
V1 was towed from the scene and V2 was driven away.
```

	asphalt	driven	...	road	V1	V2	V3	V4	west	
[	0	1	...	0	1	1	0	0	0	]

Estimer le nombre de blessés dans un rapport d'accident

Limitation de l'approche par BoW

L'approche par sac de mots

- Génère une matrice $\mathbf{X}$ de très grande dimensionnalité.

L'approche par sac de mots

- Génère une matrice $\mathbf{X}$ de très grande dimensionnalité.
- Génère des vecteurs de documents sans similitude syntaxique ou lexicale des mots.

Estimer le nombre de blessés dans un rapport d'accident

Limitation de l'approche par BoW

L'approche par sac de mots

- Génère une matrice $\mathbf{X}$ de très grande dimensionnalité.
- Génère des vecteurs de documents sans similitude syntaxique ou lexicale des mots.
- Mélange les groupes grammaticaux (i.e. l'ordre des mots ou l'appartenance d'un mot à une phrase ne sont pas conservés.)

Estimer le nombre de blessés dans un rapport d'accident

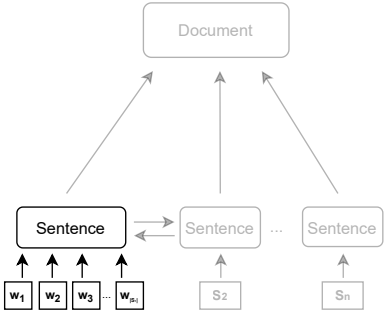
Limitation de l'approche par BoW

L'approche par sac de mots

- Génère une matrice $\mathbf{X}$ de très grande dimensionnalité.
- Génère des vecteurs de documents sans similitude syntaxique ou lexicale des mots.
- Mélange les groupes grammaticaux (i.e. l'ordre des mots ou l'appartenance d'un mot à une phrase ne sont pas conservés.)
- Ne conserve pas la structure naturelle d'un texte.

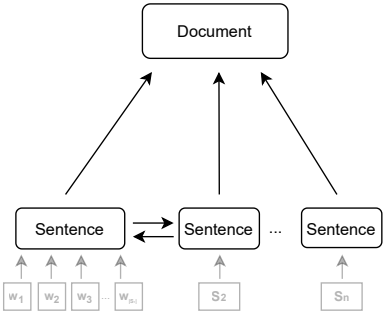
Estimer le nombre de blessés dans un rapport d'accident

Textes : données hiérarchiques en graphe



Estimer le nombre de blessés dans un rapport d'accident

Textes : données hiérarchiques en graphe



Estimer le nombre de blessés dans un rapport d'accident

Outils importants pour représenter les documents textuels

Deux outils qui ont changé la donne

Deux outils qui ont changé la donne

- Plongements lexicaux (*Word Embeddings*, p. ex. Word2Vec [Mikolov et al., 2013])
  - Réduction importante de la dimensionnalité
  - Exploitation des axes sémantiques
  - $\overrightarrow{\text{Roi}} - \overrightarrow{\text{Homme}} + \overrightarrow{\text{Femme}} = \overrightarrow{\text{Reine}}$

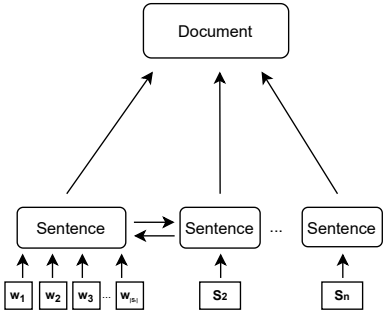
Estimer le nombre de blessés dans un rapport d'accident

Deux outils qui ont changé la donne

- Plongements lexicaux (*Word Embeddings*, p. ex. Word2Vec [Mikolov et al., 2013])
  - Réduction importante de la dimensionnalité
  - Exploitation des axes sémantiques
  - $\overrightarrow{\text{Roi}} - \overrightarrow{\text{Homme}} + \overrightarrow{\text{Femme}} = \overrightarrow{\text{Reine}}$
- Réseaux de neurones
  - Réseaux de neurones récurrents : mécanisme de mémoire interne pour conserver l'ordre des intrants.
  - Hierarchical Attention Network [Yang et al., 2016] : modélisation distincte des mots et des phrases importantes.

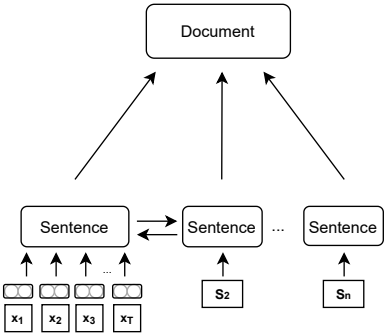
Estimer le nombre de blessés dans un rapport d'accident

Des données à HAN



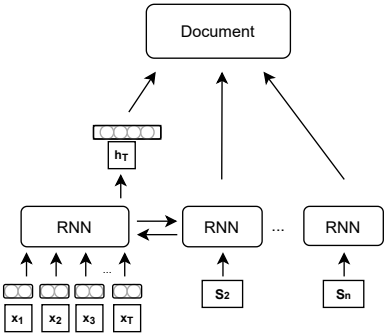
Estimer le nombre de blessés dans un rapport d'accident

Des données à HAN



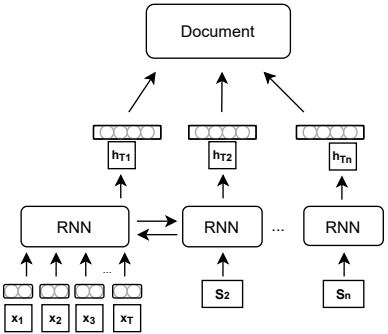
Estimer le nombre de blessés dans un rapport d'accident

Des données à HAN



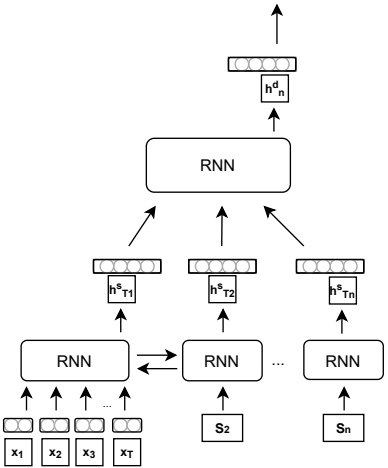
Estimer le nombre de blessés dans un rapport d'accident

Des données à HAN



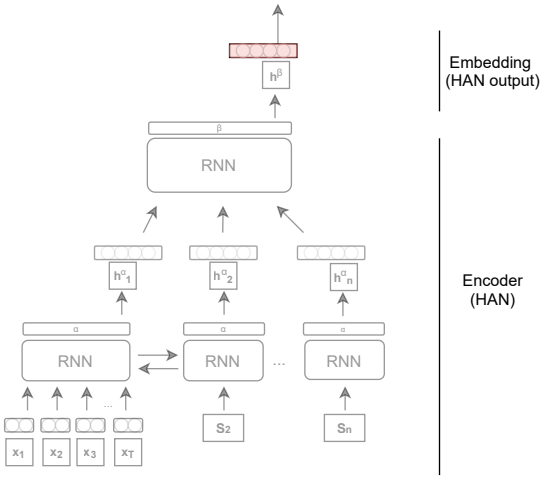
Estimer le nombre de blessés dans un rapport d'accident

Des données à HAN



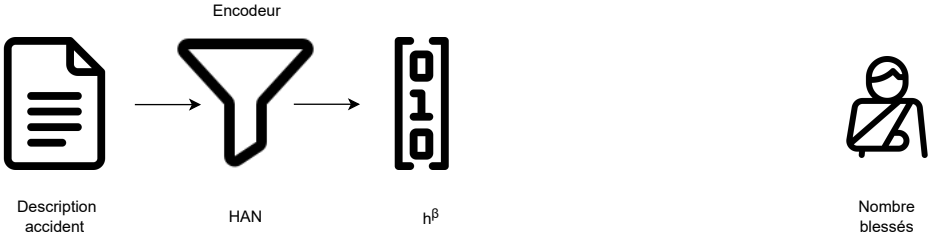
Estimer le nombre de blessés dans un rapport d'accident

Des données à HAN



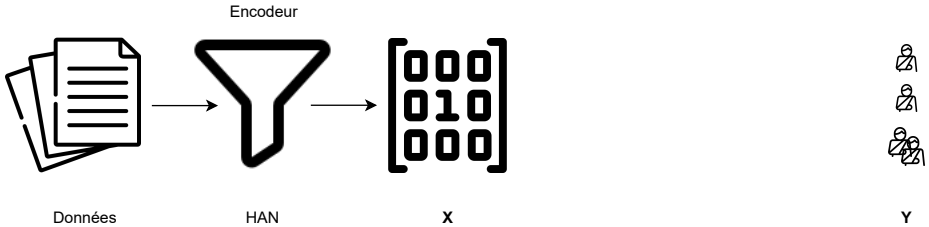
Estimer le nombre de blessés dans un rapport d'accident

Jusqu'à présent on a



Estimer le nombre de blessés dans un rapport d'accident

Jusqu'à présent on a



Estimer le nombre de blessés dans un rapport d'accident

Décodeur : sélection de la tâche

Tâche : estimer le nombre de blessés $\mathbf{Y}$

Estimer le nombre de blessés dans un rapport d'accident

Décodeur : sélection de la tâche

Tâche : estimer le nombre de blessés $\mathbf{Y}$

- Régression Poisson

Estimer le nombre de blessés dans un rapport d'accident

Décodeur : sélection de la tâche

Tâche : estimer le nombre de blessés $\mathbf{Y}$

- Régression Poisson
- Transformer les encodages en prédiction avec la fonction de lien, $\lambda = e^{h^{[\beta]}\theta'}$

Estimer le nombre de blessés dans un rapport d'accident

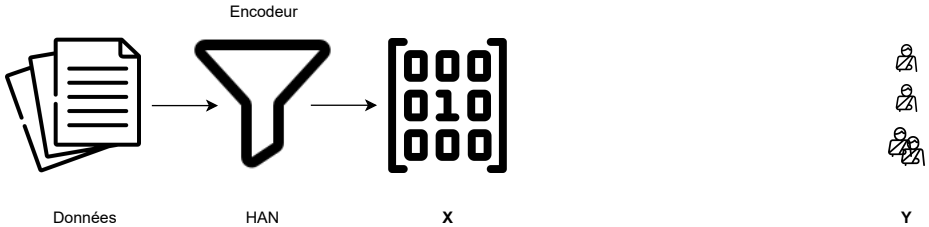
Décodeur : sélection de la tâche

Tâche : estimer le nombre de blessés $\mathbf{Y}$

- Régression Poisson
- Transformer les encodages en prédiction avec la fonction de lien, $\lambda = e^{h^{[\beta]}\theta'}$
- Minimiser log-vraisemblance : $\ell(\theta \mid X, Y) = \sum_{i=1}^{n} (y_i \log(\lambda_i) - \lambda_i)$

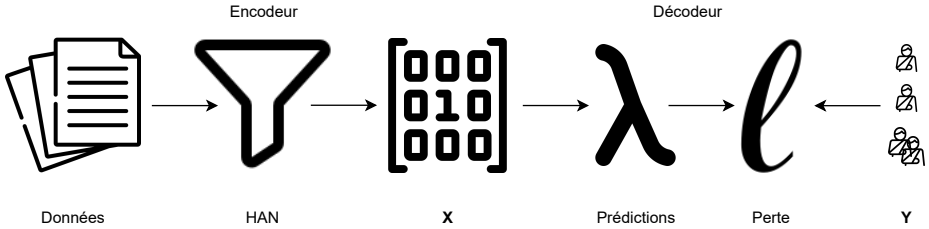
Estimer le nombre de blessés dans un rapport d'accident

Décodeur : sélection de la tâche



Estimer le nombre de blessés dans un rapport d'accident

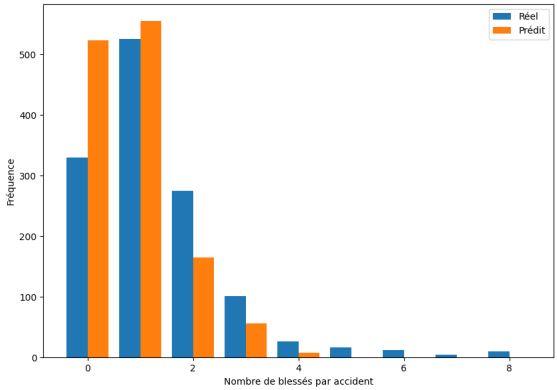
Décodeur : sélection de la tâche



Estimer le nombre de blessés dans un rapport d'accident

Résultats

Prédiction du nombre de blessés en utilisant la description du sinistre



Estimer le nombre de blessés dans un rapport d'accident

Extraire les éléments importants des séquences pour une tâche de prédiction précise

Utilisation d'un mécanisme d'explicabilité (attention [Bahdanau et al., 2015]) pour améliorer la confiance dans le modèle.

Sentence	Prediction
The driver lost control and departed the left road edge rotating counter-clockwise	1
The driver of V1 and V2 were not transported because of injuries.	0
The husband also smelled the marijuana on the driver of V1.	2
The crash occurred during the afternoon rush hour on a weekday.	0
... who reported possible injuries and were both transported to a medical facility.	4

Illustration – Poids d'attentions de la phrase la plus importante d'une description d'accident

Estimer le nombre de blessés dans un rapport d'accident

On récapitule

Dans cette étude de cas, on a

- créé un modèle explicable et portable capable d'estimer, à partir d'une description de sinistres, le nombre de personnes blessées.

Estimer le nombre de blessés dans un rapport d'accident

On récapitule

Dans cette étude de cas, on a

- créé un modèle explicable et portable capable d'estimer, à partir d'une description de sinistres, le nombre de personnes blessées.
- utilisé l'apprentissage de représentation, les plongements lexicaux et des réseaux de neurones afin de palier aux problèmes d'attributs générés manuellement dans le contexte de données non structurées.

Étude de cas : Détection de sinistres catastrophiques avec les dossiers de sinistres

Prédicteurs de sinistres catastrophiques

Mise en scène

- Les coûts des sinistres catastrophiques associés aux blessures corporelles dépassent fréquemment les montants estimés pour cause de détérioration du risque. On aimerait pouvoir détecter rapidement ces sinistres en utilisant les notes de sinistre afin de minimiser le "leakage" en intervenant plus rapidement.

Prédicteurs de sinistres catastrophiques

Mise en scène

- Les coûts des sinistres catastrophiques associés aux blessures corporelles dépassent fréquemment les montants estimés pour cause de détérioration du risque. On aimerait pouvoir détecter rapidement ces sinistres en utilisant les notes de sinistre afin de minimiser le "leakage" en intervenant plus rapidement.
- **Problème** : Les systèmes administratifs enregistrent plus de 20 000 nouveaux messages par jours.

Prédicteurs de sinistres catastrophiques

Mise en scène

- Les coûts des sinistres catastrophiques associés aux blessures corporelles dépassent fréquemment les montants estimés pour cause de détérioration du risque. On aimerait pouvoir détecter rapidement ces sinistres en utilisant les notes de sinistre afin de minimiser le "leakage" en intervenant plus rapidement.
- **Problème** : Les systèmes administratifs enregistrent plus de 20 000 nouveaux messages par jours.
- **Solution** : entraîner un modèle à reconnaître les sinistres catastrophiques avec les notes du passé, extraire les facteurs de risques et filtrer les nouvelles notes pour un analyste.

Prédicteurs de sinistres catastrophiques

Travaux reliés

- [Xu et al., 2019] utilise les séries texto-temporelles pour faire le suivi de sujet dans des bulletins de nouvelles
- [Dimri et al., 2019] et [Dimri et al., 2022] utilisent les premiers 512 mots d'un dossier de sinistre pour améliorer le traitement de la réclamation dans les processus d'affaires en assurance IARD.
- [Xu et al., 2022] détermine les coûts de garantie de remplacement, sans utiliser les données temporelles
- Travaux préliminaires indiquent que la description d'un sinistre n'est pas suffisante pour avoir une détection adéquate.

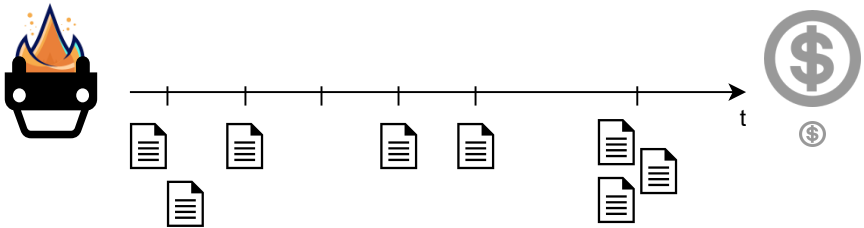
Prédicteurs de sinistres catastrophiques

Défi

- Détecter les sinistres catastrophiques.
- Modéliser automatiquement et efficacement les longues séries texto-temporelles.
- Identifier automatiquement les prédicteurs de risques (notes de sinistre) prédisant une sévérité élevée.

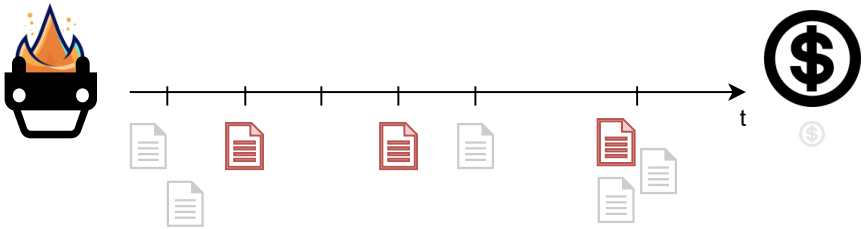
Prédicteurs de sinistres catastrophiques

On désire transformer ceci



Prédicteurs de sinistres catastrophiques

On désire transformer en cela



Prédicteurs de sinistres catastrophiques

Contenu

Reason for File Transfer : Both claimants are outside of the mig. NAME-2 was assessed by Dr LASTNAME-OTHER Day on January 28, 2016 and was diagnosed with Adjustment Disorder with Mixed Anxiety and Depressed Mood and a Somatic Symptom Disorder, Persistent, with Predominate pain. Mr NAME-1 was assessed by Dr LASTNAME-OTHER on January 20, 2016 and was diagnosed with DSM-IV criteria for a diagnosis of adjustment disorder with mixed anxiety and depression. ADR (is the file currently in mediation, arbitration, litigation and if so, at what stage ?) : No Brief File Summary : Notes on Policy coverage, Priority, WSIB, left and More

Email Sent.

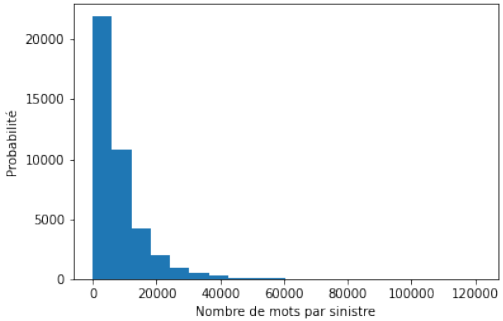
Prédicteurs de sinistres catastrophiques

Régression logistique

- $\mathbf{X}$, représentation vectorielle de l'agrégation des notes de sinistre pour les n sinistres $d_i (i \in 1, ..., n)$ survenus dans le portefeuille.
- $\mathbf{Y}$, la variable réponse (i.e. si le sinistre est catastrophique ou non)

Comment créer $\mathbf{X}$?

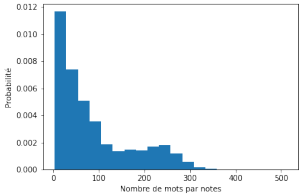
- Difficile d'utiliser des RNN pour encoder le sinistre en entier
- Séquences trop longues [Li et al., 2018] (≥ 1000)

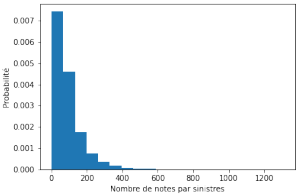


Prédicteurs de sinistres catastrophiques

Structure alternative

Mots → Notes → Dossier de réclamation





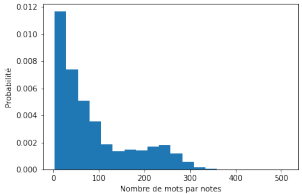
Solution Avant 2021

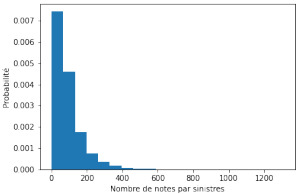
- **Mots → Notes** : RNN
- **Notes → Dossier de sinistres** : RNN

Prédicteurs de sinistres catastrophiques

Structure alternative

Mots → Notes → Dossier de réclamation

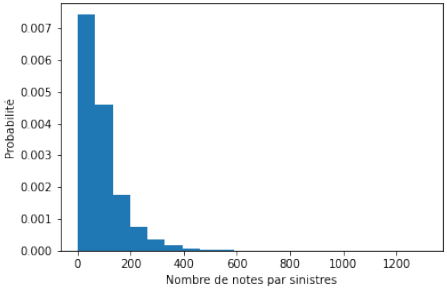


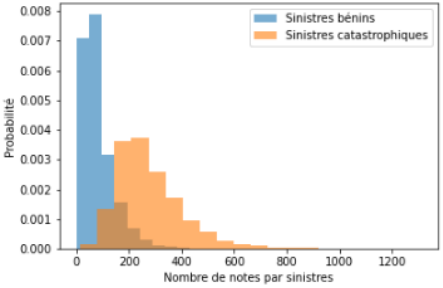


Solution Avant 2021

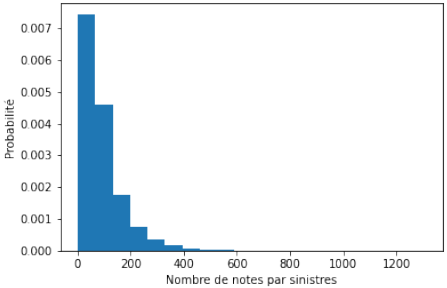
- **Mots → Notes** : RNN
- **Notes → Dossier de sinistres** : RNN
- HAN, organisé différemment

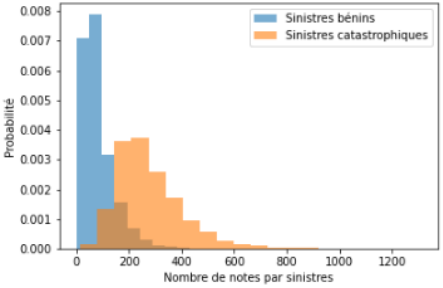
Caché dans les données





Caché dans les données





- Raccourci de classification utilisant le nombre de notes !
- Problème étudié et adressé dans [Baillargeon et al., 2022].

Solution 2023

Solution 2023

- **Mots** $\rightarrow$ **Notes** : Transformeur ([Vaswani et al., 2017]) RoBERTa ([Liu et al., 2019]) par Facebook

Solution 2023

- **Mots** $\rightarrow$ **Notes** : Transformeur ([Vaswani et al., 2017]) RoBERTa ([Liu et al., 2019]) par Facebook
  - 512 mots vs 60 000 mots
  - Limitation mémoire des modèles pour longues séquences
  - Repenser la boucle d'entraînement
  - Apprendre le vocabulaire de l'assurance au Transformeur

Prédicteurs de sinistres catastrophiques

Encodeur

Solution 2023

- **Mots** $\rightarrow$ **Notes** : Transformeur ([Vaswani et al., 2017]) RoBERTa ([Liu et al., 2019]) par Facebook
  - 512 mots vs 60 000 mots
  - Limitation mémoire des modèles pour longues séquences
  - Repenser la boucle d'entraînement
  - Apprendre le vocabulaire de l'assurance au Transformeur
- **Notes** $\rightarrow$ **Dossier de réclamation** : RNN

Solution 2023

- **Mots** $\rightarrow$ **Notes** : Transformeur ([Vaswani et al., 2017]) RoBERTa ([Liu et al., 2019]) par Facebook
  - 512 mots vs 60 000 mots
  - Limitation mémoire des modèles pour longues séquences
  - Repenser la boucle d'entraînement
  - Apprendre le vocabulaire de l'assurance au Transformeur
- **Notes** $\rightarrow$ **Dossier de réclamation** : RNN
  - Pas de Transformeur au 2e niveaux
  - Problème de raccourcis de classification (voir [Baillargeon and Lamontagne, 2023]).
- Architecture Recurrence Over Transformer, présenté par [Pappagari et al., 2019].

Prédicteurs de sinistres catastrophiques

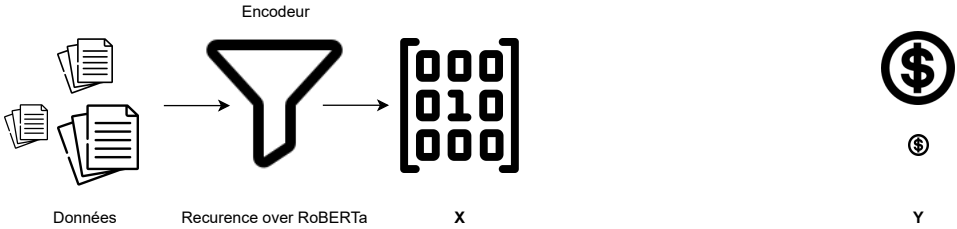
Jusqu'à présent on a

Données

Y

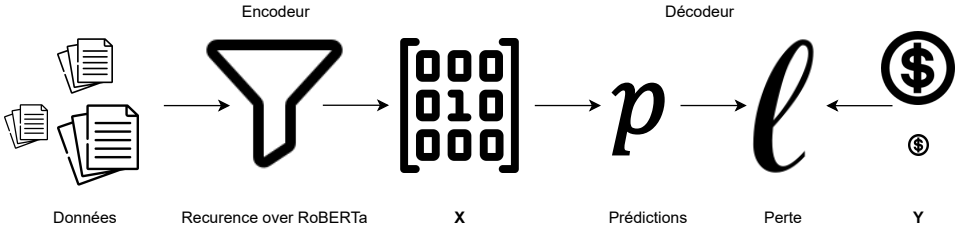
Prédicteurs de sinistres catastrophiques

Jusqu'à présent on a



Prédicteurs de sinistres catastrophiques

Jusqu'à présent on a



Prédicteurs de sinistres catastrophiques

Décodeur : sélection de la tâche

Tâche : classifier catastrophique ou non ($\mathbf{Y}$, une variable bernouilli)

Prédicteurs de sinistres catastrophiques

Décodeur : sélection de la tâche

Tâche : classifier catastrophique ou non ($\mathbf{Y}$, une variable bernouilli)

- Régression logistique

Prédicteurs de sinistres catastrophiques

Décodeur : sélection de la tâche

Tâche : classifier catastrophique ou non ($\mathbf{Y}$, une variable bernouilli)

- Régression logistique
- Transformer les encodages en prédiction, $p = \frac{1}{1+e^{-h^{[\beta]}\theta'}}$

Prédicteurs de sinistres catastrophiques

Décodeur : sélection de la tâche

Tâche : classifier catastrophique ou non ($\mathbf{Y}$, une variable bernouilli)

- Régression logistique
- Transformer les encodages en prédiction, $p = \frac{1}{1+e^{-h^{[\beta]}\theta'}}$
- Minimiser la Log-vraisemblance : $\ell(\theta \mid X, Y) = \sum_{i=1}^{n} y_i \ln p_i - (1 - y_i) \ln(1 - p_i)$

Résultats de classification des sinistres dans le jeu d'évaluation

Encodeur de notes	Précision
HAN (Phrase-Sinistres)	82.8 %
HAN (Note-Sinistres)	86.2 %
Recurrence Over RoBERTa	93.6 %

Prédicteurs de sinistres catastrophiques

Analyse de l'explicabilité du modèle

- Est-ce que les notes (prédicteurs) utilisées par le modèle sont les bonnes ?
- Est-ce que les notes tirées par notre modèle sont mieux que celles obtenues par une autre technique ?

Prédicteurs de sinistres catastrophiques

Analyse de l'explicabilité du modèle

- Est-ce que les notes (prédicteurs) utilisées par le modèle sont les bonnes ?
- Est-ce que les notes tirées par notre modèle sont mieux que celles obtenues par une autre technique ?
- Aucune méthodologie prescrite pour supporter une décision.

Prédicteurs de sinistres catastrophiques

Notre proposition

Prédicteurs de sinistres catastrophiques

Notre proposition

- Exploite la nature hiérarchique des séries texto-temporelles.

Prédicteurs de sinistres catastrophiques

Notre proposition

- Exploite la nature hiérarchique des séries texto-temporelles.
- Demande l'étiquetage manuel des notes importantes dans le sinistre.

Prédicteurs de sinistres catastrophiques

Notre proposition

- Exploite la nature hiérarchique des séries texto-temporelles.
- Demande l'étiquetage manuel des notes importantes dans le sinistre.
- Utilise un test statistique ne nécessitant aucune intervention humaine par la suite [Vilone and Longo, 2021].

Sélection aléatoire des notes importantes

- Analogie du problème des urnes des boules colorées.

Sélection aléatoire des notes importantes

- Analogie du problème des urnes des boules colorées.
- ~~Une urne~~ Un sinistre X contient N notes, dont K sont importantes

Sélection aléatoire des notes importantes

- Analogie du problème des urnes des boules colorées.
- ~~Une urne~~ Un sinistre X contient N notes, dont K sont importantes
- Si je tire n ~~boules~~ notes au hasard, quelle est la probabilité d'obtenir k notes importantes ?

Sélection aléatoire des notes importantes

- Analogie du problème des urnes des boules colorées.
- ~~Une urne~~ Un sinistre X contient N notes, dont K sont importantes
- Si je tire n ~~boules~~ notes au hasard, quelle est la probabilité d'obtenir k notes importantes ?
- Soit $C_a|X$ le nombre de notes importantes tirées aléatoirement pour un sinistre en particulier, on a que $C_a|X \sim \text{Hypergeom}(n, \frac{K}{N}, N)$.

Sélection des notes importantes par le réseau de neurones

Sélection des notes importantes par le réseau de neurones

- $C_r|X$, le nombre de notes importantes parmi les n choisies par le modèle de réseaux de neurones pour un sinistre X.

Sélection des notes importantes par le réseau de neurones

- $C_r|X$, le nombre de notes importantes parmi les n choisies par le modèle de réseaux de neurones pour un sinistre X.
- Distribution de $C_r|X$?

Sélection des notes importantes par le réseau de neurones

- $C_r|X$, le nombre de notes importantes parmi les n choisies par le modèle de réseaux de neurones pour un sinistre X.
- Distribution de $C_r|X$?
- 🙃

Sélection des notes importantes par le réseau de neurones

- $C_r|X$, le nombre de notes importantes parmi les n choisies par le modèle de réseaux de neurones pour un sinistre X.
- Distribution de $C_r|X$?
- 🙃
- On simule !

Sélection des notes importantes par le réseau de neurones

- $C_r|X$, le nombre de notes importantes parmi les n choisies par le modèle de réseaux de neurones pour un sinistre X.
- Distribution de $C_r|X$?
- 🙃
- On simule !
- Est-ce que $E[C_r|X] > E[C_a|X]$?

Résultats pour 6 réclamations

#	Nb. Notes	Nb. Imp.	$E[C_a\|X]$	$\bar{x}_{C_r\|X}$	$\hat{\sigma}_{\bar{x}_{C_r\|X}}$	p-value
1	111	30	2.70	**5.47**	0.42	0.0000
2	146	30	2.05	**3.47**	0.42	0.0010
3	143	19	1.33	**1.74**	0.26	0.0609
4	240	36	1.50	**3.44**	0.39	0.0000
5	571	95	1.66	**3.82**	0.38	0.0000
6	359	36	1.00	**1.61**	0.18	0.0007

On conclut que le modèle utilisant l'architecture Recurrence Over RoBERTa comme encodeur extrait des notes significativement plus importantes avec un seuil de significativité $(1-p)$ supérieur à 95% pour 5 des 6 cas aléatoires testés.

Prédicteurs de sinistres catastrophiques

On récapitule

Dans cette étude de cas, on a

- identifié les sinistres catastrophiques avec une précision de 93.6% en utilisant les notes de sinistres

Prédicteurs de sinistres catastrophiques

On récapitule

Dans cette étude de cas, on a

- identifié les sinistres catastrophiques avec une précision de 93.6% en utilisant les notes de sinistres
- modélisé efficacement de longues séries texto-temporelles à l'aide d'un transformeur et d'un RNN.

Prédicteurs de sinistres catastrophiques

On récapitule

Dans cette étude de cas, on a

- identifié les sinistres catastrophiques avec une précision de 93.6% en utilisant les notes de sinistres
- modélisé efficacement de longues séries texto-temporelles à l'aide d'un transformeur et d'un RNN.
- présenté une technique qui nous permet de conclure que notre modèle sélectionne mieux les prédicteurs de sinistres catastrophiques que le hasard.

Conclusion

L'exploitation des série texto-temporelles en actuariat

- Intégration de nouvelles données dans la modélisation des risques longs termes
- Modélisation des données brutes : une spécialité en soi
- Tout comme l'actuariat
- Opportunité de contribuer à deux domaines simultanément

Travaux Futurs

Évaluer la modélisation des séries texto-temporelles dans d'autres domaines

- Évaluation de l'évolution du risque de crédit en exploitant les nouvelles économiques. Similaire à [Jacobs and Hoste, 2022], mais en ajoutant l'historique des nouvelles

Modélisation explicite de la dimension *temps* des notes de sinistres

- Détection hâtive des sinistres catastrophiques
  - [Zheng et al., 2019] Calibre un modèle de survie Cox proportionnel avec les représentations des pas de temps
  - Extraire les notes plus proches temporellement de la déclaration du sinistre
- Amélioration des micro-réserves
  - [Chaoubi et al., 2022] utilise les paiements dans un LSTM
  - Intrant additionnel (paiements, probabilité de fermeture / réouverture)

Bahdanau, D., Cho, K., and Bengio, Y. (2015).
Neural machine translation by jointly learning to align and translate.
In *3rd International Conference on Learning Representations, ICLR 2015*.

Baillargeon, J.-T., Cossette, H., and Lamontagne, L. (2022).
Preventing rnn from using sequence length as a feature.
arXiv preprint arXiv:2212.08276.

Baillargeon, J.-T. and Lamontagne, L. (2023).
Reducing sequence length learning impacts on transformer models.
arXiv preprint arXiv:2212.08399.

Baillargeon, J.-T., Lamontagne, L., and Marceau, E. (2020).
Mining actuarial risk predictors in accident descriptions using recurrent neural networks.
Risks, 9(1):7.

Baker, H., Hallowell, M. R., and Tixier, A. J.-P. (2019).
Automatically learning construction injury precursors from text.

arXiv preprint arXiv:1907.11769.

Blier-Wong, C., Baillargeon, J.-T., Cossette, H., Lamontagne, L., and Marceau, E. (2021).
Rethinking representations in p&c actuarial science with deep neural networks.
arXiv preprint arXiv:2102.05784.

Borba, P. S. (2013).
Predictive analytics, text mining, and drug-impaired driving in automobile accidents.
http://us.milliman.com/.

Chaoubi, I., Besse, C., Cossette, H., and Côté, M.-P. (2022).
Micro-level reserving for general insurance claims using a long short-term memory network.
arXiv preprint arXiv:2201.13267.

Dimri, A., Paul, A., Girish, D., Lee, P., Afra, S., and Jakubowski, A. (2022).
A multi-input multi-label claims channeling system using insurance-based language models.
Expert Systems with Applications, 202:117166.

Dimri, A., Yerramilli, S., Lee, P., Afra, S., and Jakubowski, A. (2019).
Enhancing claims handling processes with insurance based language models.
In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1750–1755. IEEE.

Jacobs, G. and Hoste, V. (2022).
Sentivent : enabling supervised information extraction of company-specific events in economic and financial news.
Language Resources and Evaluation, 56(1):225–257.

Li, S., Li, W., Cook, C., Zhu, C., and Gao, Y. (2018).
Independently recurrent neural network (indrnn) : Building a longer and deeper rnn.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5457–5466.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019).
Roberta : A robustly optimized bert pretraining approach.
arXiv preprint arXiv:1907.11692.

Manski, S., Yang, K., Lee, G. Y., and Maiti, T. (2021).
Extracting information from textual descriptions for actuarial applications.
Annals of Actuarial Science, 15(3):605–622.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).
Efficient estimation of word representations in vector space.
arXiv preprint arXiv:1301.3781.

Pappagari, R., Zelasko, P., Villalba, J., Carmiel, Y., and Dehak, N. (2019).
Hierarchical transformers for long document classification.
In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Tixier, A. J.-P., Hallowell, M. R., Rajagopalan, B., and Bowman, D. (2016).
Automated content analysis for construction safety : A natural language processing system to extract precursors and outcomes from unstructured injury reports.
Automation in Construction, 62:45–56.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017).
Attention is all you need.
In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Vilone, G. and Longo, L. (2021).
Notions of explainability and evaluation approaches for explainable artificial intelligence.
Information Fusion.

Xu, G., Meng, Y., Chen, Z., Qiu, X., Wang, C., and Yao, H. (2019).
Research on topic detection and tracking for online news texts.
IEEE Access, 7:58407–58418.

Xu, S., Zhang, C., and Hong, D. (2022).
Bert-based nlp techniques for classification and severity modeling in basic warranty data study.
Insurance : Mathematics and Economics, 107:57–67.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016).
Hierarchical attention networks for document classification.
In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics : human language technologies*, pages 1480–1489.

Zappa, D., Borrelli, M., Clemente, G. P., and Savelli, N. (2019).
Text mining in insurance : From unstructured data to meaning.
Variance. In press. https://www.variancejournal.org/articlespress/.

Zheng, P., Yuan, S., and Wu, X. (2019).
Safe : A neural survival analysis model for fraud early detection.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1278–1285.