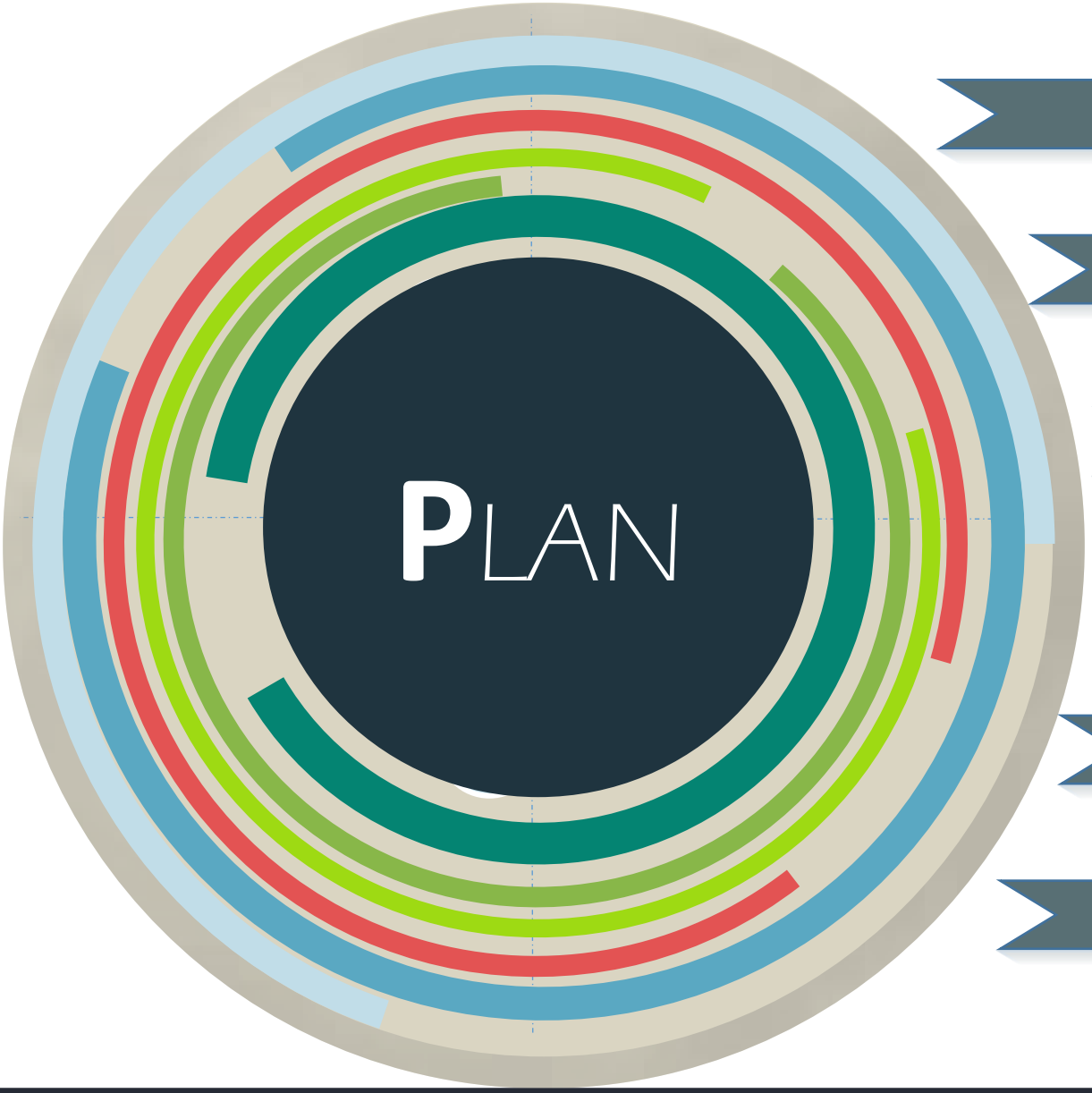
Présentation sur un article Intitulé

Neural Document Embeddings for Intensive Care Patient Mortality Prediction

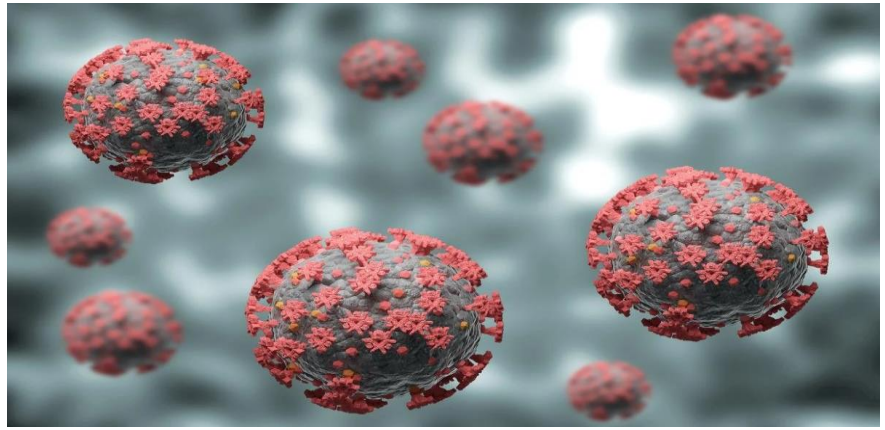
Réalisé par :

ESSALAMA Chaimae



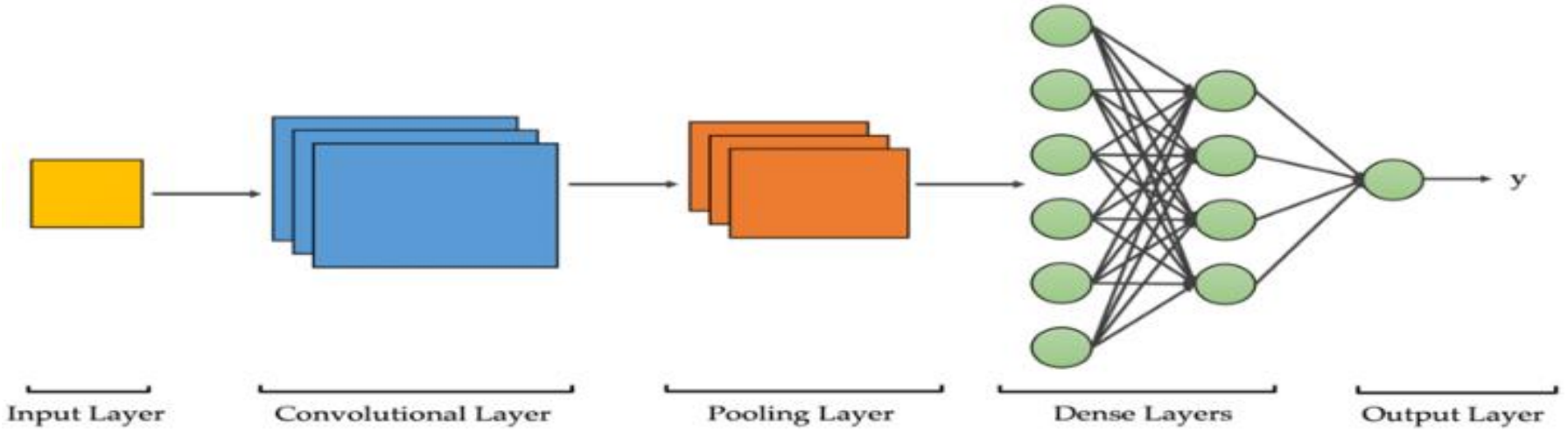
- Introduction
- Objectif
- Méthodes
- Expériences
- Résultats
- Conclusion

- Prédire avec précision les résultats futurs des patients ou du moins leur vulnérabilité et leur risque de décès.
- La prédiction de la mortalité est importante dans la pratique clinique
- la plupart des approches concurrentes reposent sur des séries chronologiques et des informations démographiques
- le traitement algorithmique de la partie textuelle non structurée des notes cliniques reste un problème important

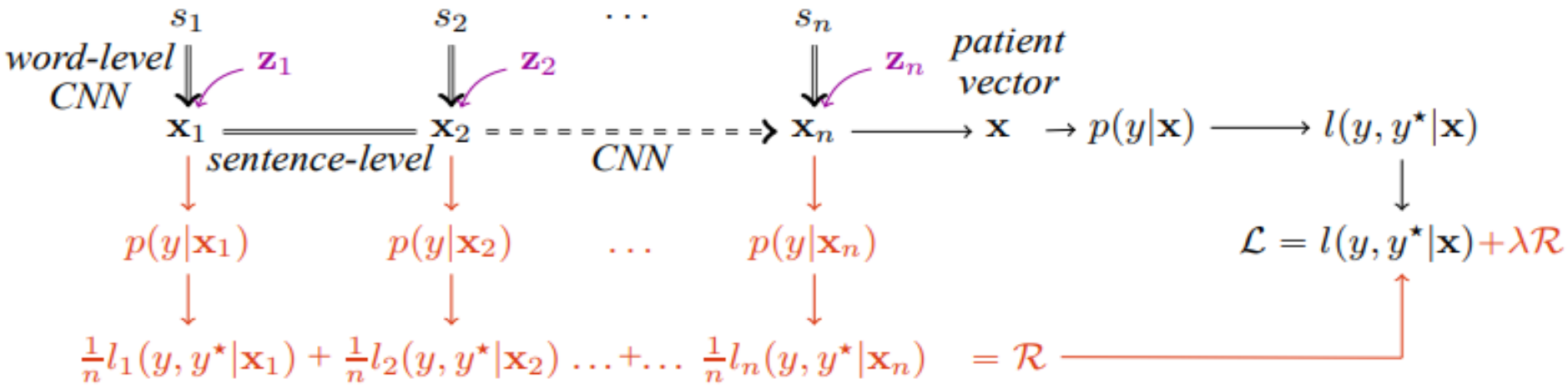


- présenter un schéma de prédiction automatique de la mortalité basé sur le contenu textuel non structuré des notes cliniques.

- Présenter une architecture de réseau neuronal convolutif qui représente explicitement non seulement des termes individuels, mais également des phrases ou des documents entiers d'une manière qui préserve ces subtilités du langage naturel.



- adopté une architecture a deux couches:
- La premier couche mappe indépendamment les phrases s_i aux vecteurs de phrase $x_i \in R^{D_s}$,
- La deuxième couche combine $< x_1, \ldots, x_n >$ dans une seule représentation du patient x $\in R^{D_p}$.
- Utiliser la convolution réseaux neuronaux (CNN) avec max-pooling
- Utiliser intégrations de mots pour fournir entrée vectorielle pour la première couche CNN
- La sortie de modèle est p(y), y $\in$ [0,1] , l'estimation probabilité de mortalité,
- l'entropie croisée l(y,y*)



Réplication cible

- Répliquer la perte aux étapes intermédiaires.
- Calculer une probabilité de mortalité softmax individuelle pour chaque phrase i = 1,…,n,
- Incorporer n termes d'entropie croisée supplémentaires dans l'objectif final
- Chercher à minimiser

$$\mathcal{L} = \sum_{(d^{(j)}, y^{*(j)}) \in \mathcal{D}} \mathcal{L}(d^{(j)}, y^{*(j)}) \qquad (1)$$

$$\mathcal{L}(d = \langle s_1, \ldots, s_n \rangle, y^*) = l(y, y^* | \mathbf{x}) + \lambda \mathcal{R} = l(y, y^* | \mathbf{x}) + \frac{\lambda}{n} \sum_{i=1}^{n} l_i(y, y^* | \mathbf{x}_i) \qquad (2)$$

Incorporer les informations de la note

- End-to-end neural network architectures permettent l'intégration facile d'informations supplémentaires pouvant augmenter le pouvoir prédictif.
- Evaluer de manière fiable l'importance des phrases individuelles pour la tâche de classification.
- Pour exploiter ces informations, nous intégrons les 14 catégories dans un espace vectoriel R^{Ds} et concaténer chaque vecteur de phrase xi avec son vecteur de catégorie associé zi.

Prédire si le patient mourra (1) pendant le séjour à l'hôpital, (2) dans les 30 jours suivant la sortie, ou (3) dans l'année suivant la sortie, et signalons l'AUC comme un mesure d'évaluation.

Données

- Utiliser les données de base de données MIMIC-III,
- limité l'étude aux adultes (18 ans et plus) avec une seule hospitalisation. Plus important encore, nous excluons les notes de la catégorie des résumés de sortie et toutes les notes enregistrées après la sortie du patient. Il en résulte 31 244 patients avec 812 158 notes
- 13,82% des patients sont décédés à l'hôpital, 3,70% sont sortis et sont décédés dans les trente jours et 12,06% sont sortis et sont décédés dans l'année
- Echantilloner au hasard 10% des patients pour l'ensemble de test et 10% pour l'ensemble de validation. Les 80% restants des patients sont utilisés pendant le training.
- Construire le vocabulaire en conservant les 300 000 mots les plus fréquents dans toutes les notes et en remplaçant tous les mots qui ne font pas partie du vocabulaire par un jeton hors vocabulaire.

Lignes de base

- tokeniser chaque note et supprimer tous les mots d'arrêt à l'aide de la liste de mots d'arrêt Onix 1
- Le vocabulaire est construit comme l'union des 500 mots les plus informatifs dans la note de chaque patient sur la base d'un tf-idf métrique.
- Tous les mots qui ne font pas partie du vocabulaire sont supprimés. Nous conservons le nombre de sujets à 50 et définissons les priorités LDA pour les distributions de sujets et les distributions sujet-mot sur:

$$\alpha = 50 / \text{number Topics} , \quad \beta = 200 / \text{vocabularySize}$$

- Nous entraînons une SVM de noyau linéaire distincte sur les distributions par sujet par note pour prédire la mortalité pour chaque tâche.
- utiliser le réseau de neurones feed-forward Comme deuxième ligne de base, en utilisant le schéma populaire de bag of words distribué (DBOW).
- former des SVM linéaires distinctes pour chaque tâche

Paramètres et pré-training

- pré-entraîner des vecteurs de mots à 50 dimensions sur les données d'entraînement à l'aide de l'implémentation word2vec de la boîte à outils gensim.
- Le CNN au niveau des mots utilise 50 filtres de tailles 3, 4 et 5, ce qui donne un représentation de phrase de la taille DS = 150.
- Intégrer des catégories dans l'espace DC =10 dimensions et utiliser 50 filtres de taille 3 pour le CNN au niveau de la phrase, ce qui donne une représentation de la taille du patient DP = 50.
- Régularisation de la couche entièrement connectée avant notre softmax final par l2-régularisation sur les poids et l'abandon avec une probabilité de maintien de 0,8.

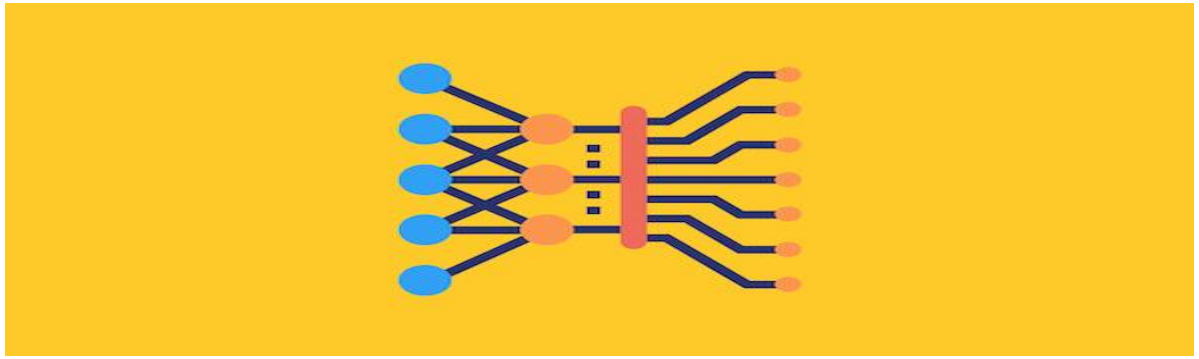


Table1: MIMIC-IIIMortality prediction AUC

Tâche	LDA	doc2vec	CNN
Hôpital	0,930	0,930	0,963
30-jours	0,800	0,831	0,858
1 an	0,790	0,824	0,853

Tableau2: Analyse des performances pour réplication cible

Modele	sans réplication de cible	avec réplication de cible
AUC	0,682	0,858

Tableau3: Les trois scoring sentences les plus élevées et les trois plus basses d'un patient dans la tâche d'un an

P (survie) élevé	**les lignes de support restantes sont inchangées. pas d'épanchement . les contours cardiomédiastinaux sont normaux.**
P(survie) faible	il s'avère maintenant qu'elle présente des lésions métastatiques dans son cerveau. impression UNK plusieurs grandes masses d'amélioration dans le cerveau avec un œdème vasogénique le plus compatible avec . amélioration des lésions dans le lobe temporal droit et le milieu droit du cerveau cohérent avec maladie métastatique .

- La prise en compte de la composition des mots et des phrases est cruciale pour identifier les modèles de texte importants.
- Ces résultats ont un impact au-delà du contexte immédiat des tâches de prédiction automatique et suggèrent des orientations prometteuses pour la recherche clinique sur l'apprentissage automatique afin de réduire la mortalité des patients.