## Article 1 : NVEILING BLACK-BOXES: EXPLAINABLE DEEP LEARNING MODELS FOR PATENT CLASSIFICATION

**Introduction :** C’est une étude sur du deep neural networks (DNN) pour classer les brevets. Le principal enjeu est que les IA de deep learning sont souvent considérées comme des boites noires où on ne comprend pas vraiment leur fonctionnement.

L’IA est censé donner la classe du brevet avec un relevance score (score de pertinence) et donner une explication du pourquoi en surlignant les mots important du texte qui lui ont permis de prédire sa classe d’appartenance afin de rendre le résultat plus explicable

On dispose pour cela de deux datasets de deux millions de brevets pour un maximum de variétés.

L’IPC (The international patent classification) donne 645 classes générales et plus de 67000 sous-groupes, ce qui rend la tâche de classification bien plus compliqué. On peut aussi noter qu’un brevet contient énormément de mots et de termes scientifiques ou de jargon propre à la communauté scientifique qui rend la compréhension beaucoup plus compliquée.

Avant des modèles de ML plus classiques étaient utilisés comme SVM (Support Vector Machine, KNN (K-Nearest-Neighbour) mais les modèles plus récents comme le DNN et RNN (Recurrent Neural Networks) et ses variants comme le CNN (Convolutional Neural Network), le LSTM (Long Short-Term Memory Networks), Bi-LSTM (Bidirectional LSTM) et GRU (Gated Recurrent Unit). La représentation brevet en format de texte est faite à l'aide de l'intégration de mots et d'un modèle pré-entraîné basé sur un transformateur.

Le problème de ces modèles DNN est que leurs résultats sont quasiment inexplicables ce qui pose des problèmes car les brevets représentent des documents légal qui doivent avoir des justifications argumentées.

L’une des idées est donc de surligner les mots importants qui ont permis au modèle de classification de prédire la classe qui permet d’augmenter la confiance des utilisateurs dans la précision du modèle. On peut d’ailleurs noter que rendre les IA capable d’expliquer le réponse constituent aujourd’hui un besoin beaucoup demander dans pleins de domaines différents.

**Fonctionnement global :** Utilisation d’un modèle uni-modal (qui ne peut prendre qu’un seum type de donnée, à la différence d’un multi-modal). Représenter les termes du brevet à l'aide d'un vecteur de caractéristiques sémantiques distribuées de grande dimension obtenu à partir d'une intégration de mots pré-entraînée des modèles puis procéder à la formation de plusieurs modèles basés sur DNN comme Bi-LSTM, CNN et CNN-BiLSTM, capables de prédire les classes (et sous classes) des brevets cité précédemment. Il faut pour finir utiliser le LRP-enabled (Layer-Wise Relevance) qui permet de surligner les mots important et donner un relevent score

**Related work :** L’apparition des transformer à beaucoup fait avancer les recherches sur ce sujet (et dans le NLP en général). On a notamment des deep bidirectional transformer (BERT), robust optimized BERT (RoBERTa), distilled BERT (DistilBERT) et XLNet pour la représentation du texte dans les tâches de NLP. Un transformer calcul des représentations de ses entrées et sorties sans utiliser de réseaux de neurones récurrents (RNN) ou de réseaux de neurones convolutifs (CNN). Cela permet aux transformateurs de gérer les données séquentielles de manière plus efficace, en capturant les dépendances et relations à long terme dans les données. Mais quand il est bidirectionnel, comme BERT, cela permet au modèle de considérer le contexte à la fois à gauche et à droite de chaque mot dans une phrase simultanément.

Ils ont fait un fine-tuning sur des modèles près-entrainé comme BERT, XLNet, RoBERTa et ELECTRA pour la représentation de texte et créer des multi-level classification pour les brevets et fait un fine-tuning sur un DistilBERT pour les sub-groups level.

**Explainable Patent Classification:** A Neuron (Sementic patent representation) -> Network Forward Flow (Using Deep Patent Modelling like Bi-LSTM, CNN, CNN-BiLSTM) -> LRP Backwards propagation flow (Explaining prediction highlighting relevant words related to a class C - Explanation generated by LRP).

Pour mettre le texte sous forme de vecteur à grande dimension, ils utilisent FastText. Les mots sont alors associés à leur contexte, ce qui est très utile, surtout pour les brevets qui ont un contexte très spécifique.

Le relevance score est calculé grâce à la back-propagation sur chacune des couches en allant de output-layer à input-layer sur pleins d’input différents pour voir comment le text est classé. Le relevance score est exprimé en pourcentage

**Experiments :** L’étude montre que dans la plupart des gros datasets utilisés, il y a un déséquilibre dans la réussite des brevets classés selon leur véritable classe d’appartenance et qui sont en plus souvent les mêmes.

On note que peut-importe le dataset utilisé, c’est le modèle de classification profond Bi-LSTM qui a le meilleur F1 score (0,69-0.78), et donc de même pour sa précision et son rappel. Même si les autres modèles (CNN et CNN-BiLSTM) ont eu des résultats assez proche.

Par rapport aux datasets utilisés, le meilleur reste BigPatent car il contient à chaque fois un résumé du brevet rédiger par des humains qui permet au modèle de comparer les données brutes et le résumé pour une meilleure compréhension.

Le classe la plus dure à déterminer est la classe générale (F1 score de 0.41) car ils ne contiennent pas assez d’information précise. C’est une classe trop vague.

A ce moment-là, leur méthode de classification a les meilleurs résultats.

Pour souligner les mots importants qui contribuent à la prédiction de classe par back-propagation, ils utilisent la couleur rouge avec un pourcentage de transparence en fonction de la pondération du mot en question.

L’un des problèmes est que certains mots peuvent être souligné alors qu’ils peuvent être commun dans plusieurs classes en même temps.

**Limitation :** Ce modèle est mauvais pour expliquer sa réponse sur les sous-classes et les sous-groupes. Ils pourraient être bien meilleur s’il avait été entrainé sur des mots spécifique à se retrouver dans un corpus de brevet et non pas des mots communs comme c’est le cas ici. Ils conseillent donc de voir du côté des transformers pour essayer de résoudre ces différents problèmes.

## Article 2: A Comprehensive Survey on AI-based Methods for Patents