Université de Montpellier

Faculté des sciences

Master 2 IPS

Résumé d'article

HMIN326M-Fouilles de données

Arezki KACIOUI
201612827
14/12/2020

Article 3 choisi : Waleed Ragheb, Jérôme Azé, Sandra Bringay, Maximilien Servajean.
"Pourquoi dois-je croire ta prédiction ?" In *30es Journées Francophones d'Ingénierie des Connaissances (IC)* (pp. 112-124).

1. Résumé de l'article

Dans cet article, les auteurs proposent un modèle utilisant l'apprentissage par transfert couplé à des mécanismes d'attention permettant l'analyse et la classification des sentiments.

La classification et l'analyse des sentiments est un domaine de recherche très actif dans la communauté de l'apprentissage automatique. Dans ce sens, plusieurs approches ont été développées et expérimentées, devenant de plus en plus performantes et efficaces. Les plus utilisées sont les words embeding et l'apprentissage par transfert souvent combiné à des réseaux de neurones. Le modèle le plus populaire dans cette dernière approche est celui proposé par Howard et al,2017, ULMFiT.

Il subsiste néanmoins quelques limites à ces modèles, notamment la dégradation des résultats lorsque la longueur de la séquence augmente, due au fait que la séquence d'entrée dans une représentation interne est d'une longueur fixe, une efficacité restreinte à certains domaines, mais aussi l'opacité de ces modèles et leurs complexité pouvant avoir un effet sur la confiance des utilisateurs.

Dans ce sens les auteurs ont pris comme baseline le model ULMFiT auquel ils ont couplé des mécanismes d'attention afin de résoudre les soucis de longueur fixe. Ils s'intéressent notamment à la sortie des mécanismes d'attention, sous forme de visualisation des parties du texte ayant impacté la prédiction et l'interprétabilité du modèle.

L'architecture choisie s'appuie sur une auto-attention basée sur un encodeur AWD-LSTM agrégée sur plusieurs niveaux de l'auto-attention, un empilement de trois Att-LSTM. Par la suite, une fois les informations agrégées, celles-ci passent par deux couches de classification qui vont déterminer la classe de sortie. Enfin, le modèle sera entrainé puis passera à l'expérimentation. Deux expérimentations majeures ont été effectuées, l'une sur les performances du modèle comparé aux principaux compétiteurs et l'autre sur son interprétabilité.

En conclusion, le modèle proposé a enregistré des résultats très compétitifs par rapport aux autres modèles, malgré une interprétation local se limitant à une prédiction unique, en plus d'impacter positivement les perceptions liées à l'interprétabilité des utilisateurs grâce au système de visualisation.

En perspective, une étude comparative quant à l'effet de la visualisation sur la confiance permettrait de confirmer cet impact. En outre, le modèle peut être amélioré afin d'avoir une interprétation plus global et une visualisation plus enrichies.

2. Questions

a. Expliquer le concept d'explicabilité des systèmes d'intelligence artificiel et en particulier des modèles d'apprentissages automatiques.

⇨ Les modèles d'apprentissage automatique sont souvent perçus comme des boites noires difficiles à interpréter. De ce fait, afin de composer avec les performances de ces modèles et la confiance qu'on leurs attribue, le concept d'explicabilité est la solution qui permet de développer des outils en mesure d'améliorer la compréhension de ces boites noir et les transformer en boites grises, dont le processus de prise de décision est partiellement décomposé et expliqué.

b. Le positionnement des auteurs par rapport à l'état de l'art dans le domaine étudié :

- ⇨ Les auteurs ont pris comme baseline pour leurs travaux, des travaux déjà existants notamment (Howard & Ruder, 2018) qui ont proposé la méthode ULMFiT basée sur une représentation des mots peu profondes, et (Merity et al., 2018) qui ont combiné la méthode ULMFiT à un apprentissage par transfert.
- ⇨ Il existe aussi une autre approche basé sur les words embedings.
- ⇨ **L'approche proposée est différente par rapport à l'existant :**
- ⇨ Les auteurs ont utilisé l'architecture ULMFiT (éxistant) intégrant un apprentissage par transfert et y ont intégré des mécanismes d'attention pour l'analyse de sentiments afin d'évaluer dans quelle mesure une visualisation basée sur le mécanisme d'attention facilite la prise de décision.

c. La différence entre les deux types d'expériences qui ont été faites :

- ⇨ **Les auteurs ont effectué deux expérimentations avec le modèle développé :**
  - 1. Expérimentations sur les performances du modèle
  - 2. Expérimentations sur l'interprétabilité du modèle
  - Les deux expérimentations ont un but différent, l'une voulant étudier la performance du modèle, l'autre son interprétabilité et son explicabilité.
- ⇨ **Chaque test veut démontrer :**
  - 1. Expérimentations sur les performances du modèle : Veut démontrer les performances du modèle sur des jeux de données et le comparer à d'autres modèles existants.
  - 2. Expérimentations sur l'interprétabilité du modèle : Veut démontrer l'interprétabilité du modèle et son explicabilité basé sur les mécanismes de l'attention.

d. Donner trois faiblesses et trois points forts de l'article :

- ⇨ **Points forts**
  - Modèle expressif offrant une représentation visuelle permettant de structurer les explications.
  - Modèle plus performant que les modèles classiques d'apprentissage basé sur les words embeddings et aussi performant que les modèles à la pointe de l'état de l'art.
  - L'une des premières études à s'intéresser à l'aspect qualitatif de l'impact de la visualisation de l'attention sur l'interprétation que les utilisateurs peuvent faire de cette nouvelle information dans le cas de la classification de données textuelles et intérêt majeur à l'explicabilité du modèle et la confiance des utilisateurs.
- ⇨ **Faiblesses**
  - Interprétation local se limitant à une prédiction unique.
  - Peu d'expérimentations, et tests du modèle sur seulement cinq jeux de données.
  - Manque d'études comparatives quant à l'effet de la visualisation et la confiance qu'un utilisateur a en un modèle.