ChouBERT : Pré-entraînement d'un modèle de langue française pour le Crowdsensing avec des Tweets dans un contexte phytosanitaire

Shufan Jiang, Rafael Angarita, Stéphane Cormier, Julien Orensanz, Francis Rousseaux

ChouBERT : Pré-entraînement d'un modèle de langue française pour le Crowdsensing avec des Tweets dans un contexte phytosanitaire

Shufan Jiang [1,2], **Rafael Angarita** [3,4], **Stéphane Cormier** [2], **Julien Orensanz** [5], **Francis Rousseaux** [2]

1. Valda, DIENS, École normale supérieure, Paris, France
jiang.chou.fan@gmail.com

2. CRESTIC, Université de Reims Champagne Ardenne, Reims, France

3. Université Paris Nanterre, Nanterre, France

4. LIP6, Sorbonne Université, Paris, France

5. Cap2020, Gradignan, France

RÉFÉFRENCE DE L'ARTICLE INTERNATIONAL. *ChouBERT: Pre-training French Language Model for Crowdsensing with Tweets in Phytosanitary Context. RCIS 2022: 653-661.*

MOTS-CLÉS : *apprentissage par transfert, détection des foules, surveillance de la santé végétale, twitter*

KEYWORDS: *transfer learning, crowd-sensing, plant health monitoring, twitter*

Crowdsensing ou la *détection des foules* est un paradigme de détection qui permet aux personnes ordinaires de contribuer avec des données détectées ou générées par leurs appareils mobiles équipés de capteurs (Boubiche *et al*., 2019). Il introduit un nouveau changement dans la manière dont nous collectons les données en permettant d'acquérir des connaissances locales par le biais de dispositifs intelligents portés par les gens, tels que les smartphones, les tablettes, les montres intelligentes, entre autres. Cela permet d'exploiter les capteurs améliorés des smartphones de manière rapide et économique, contrairement aux méthodes traditionnelles plus coûteuses. Poussés par la reconnaissance croissante de l'importance de l'agriculture pour le maintien de l'humanité et le rôle central des agriculteurs dans la transformation numérique de l'agriculture , nous avons témoigné de l'émergence d'applications de crowdsensing pour l'agriculture (Mendes *et al*., 2020).

Les agriculteurs sont également de plus en plus présents dans les médias sociaux tels que Facebook, WhatsApp et Twitter, où ils partagent et discutent volontairement

leurs observations sur l'environnement. En particulier, Twitter permet aux agriculteurs de publier librement de courts messages appelés "tweets" pour partager leurs observations. Pour tirer parti de ces observations, il faut garder la trace des sources de données pertinentes parmi le bruit, extraire et organiser les informations qu'elles contiennent et les partager avec d'autres utilisateurs intéressés. Cela n'est possible qu'au prix d'un effort humain important, en inspectant, filtrant et nettoyant manuellement toutes les données et en reliant les entités et les contextes connexes.

Les applications récentes de modèle de langage semblent prometteuses pour résoudre les problèmes d'extraction d'informations. Poussés par la connectivité croissante des agriculteurs et l'émergence de communautés agricoles en ligne, notre objectif est d'explorer l'application émergente des observations à la ferme via les réseaux sociaux -en particulier Twitter- et de proposer une approche pour la classification des tweets. Nous cherchons à répondre aux questions de recherche suivantes : *RQ1. comment des modèles de langage (LM) pré-entraînés peuvent aider à l'exploration des tweets sur la snta végétale ?* ; et *RQ2. comment adapter les LM génériques pour la classification de textes dans un domaine spécifique ?*

Nous avons construit ChouBERT [1] en pré-entraînant CamemBERT (Martin *et al.*, 2020) sur les bulletins de santé des végétaux (BSV) et les tweets pour augmenter l'intégration contextuelle des tweets pour la détection des observations. Nos résultats expérimentaux mettent en évidence la généralisation de la représentation de ChouBERT sur des aléas non vus pour la tâche de classification. Nous pouvons généraliser cette approche pour améliorer le crowdsensing basé sur le contenu textuel des tweets : en collectant un jeu initial de tweets à l'aide de mots-clés; en étiquetant manuellement un petit jeu de tweets; en pré-entraînant davantage les modèles de langage à l'aide de documents du domaine et de tweets; et en construisant des applications NLP avec le jeu étiqueté et le modèle de langage adapté au domaine. Enfin, notre expérience montre que l'observation des foules sur Twitter ne remplace pas les autres paradigmes de surveillance, mais constitue une source d'information complémentaire. L'objectif du crowdsensing sur Twitter est de détecter les signaux faibles plutôt que de quantifier la gravité d'un problème par la fréquence des mentions. Il peut être intéressant de croiser ces informations avec d'autres sources de données.

Bibliographie

Boubiche D. E. *et al.* (2019). Mobile crowd sensing–taxonomy, applications, challenges, and solutions. *Computers in Human Behavior*, vol. 101, p. 352–370.

Martin L. *et al.* (2020). Camembert: a tasty french language model. In *Proc. of the 58th annual meeting of the association for computational linguistics*, p. 7203–7219. Online, Association for Computational Linguistics.

Mendes J. *et al.* (2020). Smartphone applications targeting precision agriculture practices—a systematic review. *Agronomy*, vol. 10, n° 6, p. 855.

1. https://huggingface.co/ChouBERT