Groupe 6 : évaluation du caractère 'fake news' de messages sur les réseaux sociaux

Arthur BAUDOT
baua1101

Victor MANTEL
manv0801

Raphaël LION
lior0601

1 Introduction

2 État de l'art

Ce problème peut se ramener à une tâche de classification binaire, dans laquelle la classe positive indique que le tweet vaut la peine d'être vérifié. Pour prioriser les tweets à vérifier, la sortie du modèle de classification (c'est-à-dire la probabilité du tweet d'être dans la classe positive) donne le score de priorité du tweet (Jaradat et al., 2018).

2.1 Sélection et extraction des caractéristiques

S'il est important d'utiliser un bon modèle de classification, l'essentiel du travail réside dans la sélection et l'extraction des caractéristiques du tweet. Ceci commence par sélectionner des types de caractéristiques qui permettraient de distinguer plus facilement une fausse information d'une vraie information. Les types de caractéristiques les plus utilisés sont le score de sentiment, la longueur du tweet, l'embedding des mots du tweet (avec une représentation *tf-idf* ou *word2vec* (Mikolov et al., 2013)), la valeur grammaticale de chaque mot (Parts Of Speech (POS) tag), les types d'entités présentes dans le tweet (Hassan et al., 2015) (Gencheva et al., 2017). La plupart de ces caractéristiques peuvent être déterminées à l'aide de l'API *AlchemyAPI* et de la librairie *NLTK*. Cependant, ces types de caractéristiques considèrent les phrases du tweet indépendantes, et il peut être pertinent d'inclure des caractéristiques prenant en compte le contexte des phrases. On peut notamment utiliser la position de la phrase dans le tweet, les tailles des phrases précédentes et suivantes, déterminer le sujet du tweet, ou encore identifier des contradictions (Gencheva et al., 2017).

Les caractéristiques énoncées ci-dessus sont des types de caractéristiques, mais le nombre réel de caractéristiques dépend des tweets considérés (le nombre d'entités détectées, le nombre de POS tags...). Pour éviter d'avoir trop de caractéristiques en entrée du modèle de classification, il est nécessaire de sélectionner les caractéristiques les plus discriminantes. Pour ce faire, on peut notamment entraîner un modèle de forêts aléatoires en mesurant l'impureté d'une caractéristique (Hassan et al., 2015). Plus l'impureté d'une caractéristique est faible, plus celle-ci est importante car elle détermine mieux la décision finale.

2.2 Modèles de classification

Les modèles de classification supervisée utilisés varient beaucoup selon les études. Beaucoup d'études essaient des modèles de classification linéaire comme les SVM (Hassan et al., 2015) (Zuo et al., 2018) (Ghanem et al., 2018) pour effectuer la prédiction. D'autres études utilisent des méthodes non linéaires comme des réseaux de neurones (Gencheva et al., 2017) (Jaradat et al., 2018) (Hansen et al., 2018) ou des méthodes à noyaux (Gencheva et al., 2017). Pour tous les articles étudiés, une 4-fold cross validation a été utilisée pour estimer les meilleurs hyper-paramètres à utiliser.

2.3 Mesures d'évaluation

Les mesures d'évaluation les plus utilisées par les chercheurs pour ce problème sont la *Précision à k (P@k)* et la *Mean Average Precision (MAP)*. La mesure $P@k$ représente la précision du modèle en évaluant celle-ci sur k tweets. La *MAP* est la moyenne de l'*Average Precision* sur les jeux de données utilisés, sachant que l'*Average Precision* représente la moyenne des $P@k$ sur toutes les valeurs de k testées (StackOverFlow).

2.4 Points forts et limites

Les méthodes précédentes permettent d'extraire des caractéristiques pertinentes pour la classification, en analysant la grammaire et le contexte des

phrases, et en identifiant des propriétés caractéristiques de fausses informations.

Cependant, toutes les études précédentes travaillaient sur des débats et des discours politiques. Or, dans le cadre des tweets, il est plus difficile d'identifier le contexte car un tweet contient peu de phrases. Il est donc assez probable que les méthodes ci-dessus fonctionnent moins bien sur des tweets, et cela nous forcerait à nous contenter de considérer les tweets de façon indépendante.

Pour contrer ce problème, nous pourrions regrouper les tweets par thème et par situation, et ajouter des caractéristiques de comparaison entre eux, comme déterminer si l'information d'un tweet est en accord avec l'information générale identifiée par les autres tweets. De ce fait, nous pourrions identifier des tweets très divergents des autres, ce qui serait un indice de fausse information.

3 Méthodologie

D'abord reproduire l'architecture de base, ie choisir les types de features les + courants, faire un selectFromModel avec scikit learn pour sélectionner les features les + discriminants, et avoir un modèle de classification (en tester plusieurs).

Puis ajouter des features + adaptés à notre contexte de tweets, avec par ex comparer les tweets entre eux, etc.

4 Expériences

5 Conclusion

References

Pepa Gencheva, Preslav Nakov, Lluis Marquez, Alberto Barrón-Cedeno, and Ivan Koychev. 2017. A context-aware approach for detecting worth-checking claims in political debates.

Bilal Ghanem, Manuel Montes y Gómez, Francisco Rangel, and Paolo Rosso. 2018. Upv-inaoe-autoritas - check that: Preliminary approach for checking worthiness of claims.

Casper Hansen, Christian Hansen, Jakob Grue Simonsen, and Christina Lioma. 2018. The copenhagen team participation in the check-worthiness task of the competition of automatic identification and verification of claims in political debates of the clef-2018 checkthat! lab.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. Claimrank: Detecting check-worthy claims in arabic and english.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

StackOverFlow. https://stackoverflow.com/questions/55748792/understanding-precisionk-apk-mapk.

Chaoyuan Zuo, Ayla Ida Karakas, , and Ritwik Banerjee. 2018. A hybrid recognition system for check-worthy claims using heuristics and supervised learning.