HOURDIN Charlène | Novembre 2022

DÉTECTEZ LES BAD BUZZ

GERER SON E-REPUTATION GRACE A L'INTELLIGENCE ARTIFICIELLE

Internet a créé l'identité numérique de chacun, ce qui a conduit à une prise de conscience qu'elle peut être un avantage ou un inconvénient dans la construction d'une réputation et d'une image en ligne.
Les **Bad Buzz** peuvent avoir de graves conséquences sur une entreprise, notamment sur sa crédibilité, ses résultats commerciaux, et même la confiance qu'elle a avec ses propres employés.

Les grands groupes ont compris l'importance de la gestion de **l'e-réputation** et ont mis en place des cellules d'expertises pour surveiller les publications sur le web. Cela leur permet d'estimer la visibilité d'un sujet et de réagir en cas de mention négative.

Dans cet article, je vais vous présenter trois approches permettant d'analyser les sentiments afin de repérer les premiers signaux d'une crise potentielle.

Modèle sur mesure simple	**Modèle sur mesure avancé**	**Modèle avancé BERT**
Modèle classique (ex : régression logistique)	Modèle basé sur des réseaux de neurones profonds	Modèle de Transformers bi-directionnel pré-entraînés

Pour cette étude, nous avons utilisé le jeu de données **Sentiment140** contenant 1,6 million de tweets, pouvant être utilisés pour détecter le sentiment.

Les tweets ont été annotés :
- 0 = négatif,
- 1 = positif

Un échantillon adapté a été utilisé pour chacune des approches utilisées.

Code source de l'API: **https://github.com/charlenehourdin/analyse-de-sentiment**
Test de l'API : **https://analyse-de-sentiment.herokuapp.com/**

MÉTHODOLOGIE

Il y a deux étapes importantes à suivre lorsque nous modélisons une langue à partir de sources textuelles:

- **le prétraitement**, qui normalise le texte
- **la vectorisation**, qui transforme le texte en nombres pour que la machine puisse l'utiliser.

Dans notre prétraitement, nous avons supprimé les majuscules, les ponctuations, les caractères spéciaux et les nombres.

Afin de réduire la taille du vocabulaire, nous avons remplacé certains contenus spécifiques aux tweets *(les mentions d'utilisateurs, les liens hypertextes et les mots-dièse par user, url et hashtag)*.

Pour la vectorisation, l'approche *Modèle sur mesure avancée* utilise le plongement de mots (voir encadré).

MODÉLISATION

- **Simple : Regression logistique** : Simple et rapide à exécuter, c'est une méthode de classification linéaire qui apprend la probabilité qu'un échantillon appartienne à une certaine classe. Elle est souvent utilisée pour la classification et l'analyse prédictive. La régression logistique estime la probabilité qu'un événement se produise, tel que voter ou ne pas voter, sur la base d'un ensemble de données de variables indépendantes.
- **Avancé : LSTM avec Glove**: La couche LSTM est une couche qui possède une mémoire interne appelée cellule. La cellule permet de maintenir un état aussi longtemps que nécessaire. Cette cellule consiste en une valeur numérique que le réseau peut piloter en fonction des situations en y ajoutant une étape de plongement de mots, nous avons obtenu de bons résultats.
- **BERT :** C'est un modèle faisant partie de la classe des Transformers *(Un transformer est un modèle qui fonctionne en effectuant un petit nombre constant d'étapes. À chaque étape, il applique un mécanisme d'attention pour comprendre les relations entre les mots de la phrase, quelles que soient leurs positions respectives.)* de la bibliothèque HuggingFace. Ce modèle offre de très bonnes performances dans le traitement automatique de la langue.

WORD - EMBEDDING (plongement de mots)

Les incorporations de mots sont des représentations vectorielles de mots où les mots ayant une signification similaire ont une représentation similaire.
Les vecteurs de mots sont l'un des moyens les plus efficaces de représenter des mots.

Il est capable en réduisant la dimension de capturer le contexte, la similarité sémantique et syntaxique *(genre, synonymes, ...)* d'un mot.

Il existe plusieurs modèles, Les plus connus sont :

- **Word2vec :** c'est une technique pour apprendre les representations vectorielles des mots. Elle est basée sur l'hypothese selon laquelle les mots qui ont des voisins sémantiquement similaires ont eux-memes un sens similaire.
- GlOVE *(Global Vectors for Word Representation)* : c'est un modèle de représentation des mots basé sur les cooccurrences globales de mots dans un corpus. Il utilise l'erreur quadratique moyenne comme fonction de perte et peut préserver les relations et les similitudes entre les mots.

COMPARAISON DES MODÈLES

	Modèle simple (Regression logistique)	**Modèle avancé** (Glove lem)	**Modèle BERT**
F1-score	0,82	**0,81**	0,67
AUC	0,82	**0,90**	0,75
Precision	0,82	**0,82**	0,67
Temps (s) (prediction)	0.05s	**1,57s**	23,15s

- Le **modèle simple** est plus rapide à exécuter et offre de bonnes performances.
- Le **modèle avancé** et le **modèle BERT** nécessite des profils technique avancée et une expertise avérée
- Les **modèles de réseaux de neurones** profonds sont plus coûteux en coût et en temps d'apprentissage.

Chaque méthode a ses avantages comme ses inconvénients.

Le choix d'une approche dépendra des besoins exprimés par les équipes de l'entreprise et des ressources disponibles