Institut National de Statistique et d'Économie Appliquée (INSEA)

Rabat — Maroc

Extraction automatique des informations clés dans les annonces d'emploi à l'aide de modèles pré-entraînés

---

Projet Encadré par :
Prof. JANATI

Projet Réalisé par :
El MOUBACHOUR Oumaima
El AZAOUI Maroua
HALLA Hajar
LAHRACH Nouhaila

Année universitaire : 2025–2026

Résumé

Les plateformes de recrutement en ligne génèrent un volume important d'annonces, souvent longues et hétérogènes. Dans ce contexte, l'extraction automatique d'informations clés (intitulé du poste, localisation, entreprise, rémunération, type de contrat, exigences, avantages) est essentielle pour améliorer la recherche, la comparaison d'offres et l'analyse des risques. Ce projet propose une chaîne de traitement fondée sur des modèles pré-entraînés (Transformers) pour identifier les entités pertinentes, structurer les champs et produire une sortie normalisée. En complément, et afin de sécuriser l'usage, un module de détection d'annonces frauduleuses est étudié à partir du contenu textuel. Les résultats montrent que des approches classiques (TF–IDF + réduction de dimension + KNN/SVM) restent très compétitives sur la détection de fraude, tandis que les embeddings pré-entraînés apportent une compréhension sémantique utile, notamment pour l'extraction et la robustesse linguistique.

Mots-clés : Information Extraction, NER, Transformers, SBERT, TF–IDF, SVD, Détection de fraude.

Table des matières

1 Introduction

Avec la multiplication des plateformes de recrutement en ligne, les annonces d'emploi frauduleuses représentent un risque croissant pour les candidats. Ces annonces peuvent entraîner des pertes financières, des vols de données personnelles ou des abus de confiance. L'objectif de ce projet est de concevoir un système automatique capable de classifier les annonces d'emploi en "frauduleuses" ou "non frauduleuses", en se basant exclusivement sur leur contenu textuel.

Le projet s'inscrit dans une démarche progressive : partir de modèles simples et interprétables basés sur des représentations fréquentielles, puis évoluer vers des modèles pré-entraînés capables de capturer la sémantique du texte, afin d'évaluer les gains réels en performance.

2 Nettoyage du jeu de données

Le jeu de données initial contient 17 876 annonces. Après analyse, seules 10 000 annonces ont été conservées. Cette réduction s'explique par :

la suppression des annonces dont les champs textuels étaient vides ou quasi inexistants,

l'élimination des observations avec des valeurs manquantes sur la variable cible fraudulent,

la conservation d'un sous-ensemble cohérent garantissant une qualité textuelle suffisante pour l'analyse.

Après nettoyage, la proportion d'annonces frauduleuses est d'environ 4,67 %, ce qui met en évidence un fort déséquilibre de classes, un aspect central dans l'interprétation des performances des modèles.

3 Structuration du texte et définition des variables

Chaque annonce est décrite par plusieurs champs textuels : titre, localisation, profil de l'entreprise, description, exigences et avantages.

Afin d'exploiter l'information de manière globale, ces champs ont été concaténés pour former un document textuel unique par annonce. La variable explicative principale est donc le texte complet de l'annonce, tandis que la variable cible est :

fraudulent = 1 : annonce frauduleuse

fraudulent = 0 : annonce légitime

Cette structuration permet aux modèles de capturer des indices répartis sur l'ensemble de l'annonce (ton du texte, promesses excessives, absence d'informations précises, etc.).

4 Séparation apprentissage / test

Les données ont été séparées en deux ensembles :

80 % pour l'apprentissage

20 % pour le test

La séparation a été réalisée de manière stratifiée, afin de préserver la proportion d'annonces frauduleuses dans chaque ensemble. Cette précaution est essentielle dans un contexte

de classes déséquilibrées, pour garantir une évaluation réaliste des performances.

5 Vectorisation TF-IDF

La première représentation du texte repose sur la méthode TF-IDF (Term Frequency – Inverse Document Frequency). Cette approche transforme chaque annonce en un vecteur numérique qui reflète l'importance relative des mots :

les mots très fréquents dans tout le corpus (ex. the, and) sont pénalisés,

les mots spécifiques à certaines annonces sont mis en valeur.

La matrice TF-IDF obtenue est de très grande dimension (plus de 80 000 termes), ce qui rend nécessaire une réduction de dimension avant l'application des modèles.

6 Réduction de dimension par PCA (Truncated SVD)

Une Truncated SVD (équivalent d'une ACP pour données creuses) a été appliquée afin de projeter les données dans un espace de 300 composantes.

Les résultats montrent que :

300 composantes expliquent environ 42,5 % de la variance totale,

le seuil de 80 % de variance n'est pas atteint, ce qui est courant en analyse textuelle à haute dimension.

Cette étape permet néanmoins de réduire le bruit, de faciliter l'apprentissage et d'améliorer la stabilité des modèles.

7 Modèle SVM linéaire — résultats et interprétation

Un SVM linéaire, ajusté par validation croisée, a été appliqué sur les données TF-IDF réduites.

Résultats principaux

F1-score moyen en validation croisée 0,53

Accuracy sur le test 93,8 %

Recall pour la classe frauduleuse 86 %

Interprétation

Le SVM parvient à détecter une grande partie des annonces frauduleuses (bon recall), ce qui est crucial dans un contexte de fraude. Cependant, la précision sur la classe frauduleuse reste limitée, ce qui signifie que certaines annonces légitimes sont classées à tort comme frauduleuses.

Ce comportement est typique d'un modèle linéaire confronté à un langage varié et parfois ambigu.

8 KNN sur TF-IDF + PCA — choix du meilleur K

Le modèle K-Nearest Neighbors a ensuite été testé pour différentes valeurs de K.

Résultats observés

Le meilleur compromis est obtenu pour K = 5

F1-score 0,77 pour la classe frauduleuse

Accuracy 98 %

Interprétation

Le KNN exploite efficacement la proximité entre annonces similaires dans l'espace réduit. Il surpasse le SVM sur le F1-score, indiquant une meilleure balance entre précision et rappel pour la détection des fraudes. Cela suggère que les annonces frauduleuses forment des groupes cohérents dans l'espace TF-IDF.

9 Embeddings pré-entraînés BERT (SBERT)

Afin de dépasser les limites des approches fréquentielles, des embeddings pré-entraînés SBERT ont été utilisés.

Chaque annonce est :

découpée en segments,

encodée par SBERT,

représentée par la moyenne des embeddings des segments.

Cette approche permet de capturer la sémantique globale du texte, indépendamment de la fréquence exacte des mots.

10 KNN sur embeddings SBERT — résultats

Le KNN appliqué sur les embeddings SBERT donne :

Meilleur K = 1

F1-score 0,67

Accuracy 97 %

Interprétation

Les embeddings SBERT permettent une meilleure compréhension sémantique des annonces, notamment lorsque la fraude repose sur le sens implicite du discours. Toutefois, dans ce projet, les performances restent légèrement inférieures à celles du KNN sur TF-IDF, probablement en raison :

du faible nombre d'exemples frauduleux,

du coût computationnel élevé limitant l'apprentissage sur l'ensemble des données.

11 Analyse qualitative — affichage des annonces

```
Aperçu résultats (5 lignes):

                                                title              location  \
7240                                    Product Manager  US, PA, Philadelphia
3010  Non-Urgent Patient Transfer Attendant - Kitchener     CA, ON, Kitchener
6027                                PHP/LAMP Developer       GB, GBN, London
2040  PR4 2AS Business Admin Apprenticeship availabl...         GB, , Preston
2469                    Back Office Junior PHP Developer       EE, 37, Tallinn

                                       company_profile  \
7240                                               NaN
3010  Voyageur is one of Ontario's leading transport...
6027  MarketInvoice is one of the most high-profile ...
2040  Established on the principles that full time e...
2469  Our CompanyAdcash® is an international adverti...

                                           description  \
7240  Curalate is looking for a passionate and exper...
3010  Voyageur Medical Transportation is the largest...
6027  We're looking for an outstanding PHP developer...
2040  This is fantastic opportunity for someone want...
2469  Our Back Office Junior PHP Developer should be...


                                            requirements  \
7240  Responsibilities:Work with founders to documen...
3010  Valid Emergency Care/First Responder or Emerge...
6027  Experience building full-stack applications, u...
2040  Government funding is only available for 16-18...
2469  Excellent knowledge in PHP and MySQLGood motiv...

                                              benefits  fraudulent  \
7240  Curalate is the world's leading marketing and ...           0
3010  Full time and part time positions available. C...           0
6027  Full time salary of £40-60,000 depending on ex...           0
2040                                  Future Prospects            0
2469  Friendly atmosphereHighly competitive salarySo...           0
```

L'affichage des premières lignes du jeu de test montre que le modèle SBERT est capable de produire des prédictions cohérentes en tenant compte du contenu global (titre, description, exigences, avantages). Certaines annonces paraissant légitimes sont correctement classées comme non frauduleuses, ce qui confirme la pertinence de l'approche sémantique.

12 Limites et Perspective

12.1 Limites

— **Longueur des textes** : nécessité de segmentation (risque de perdre des dépendances lointaines).

— **Étiquettes faibles** : les champs existants sont utiles, mais ne constituent pas toujours une vérité terrain parfaite.

— **Déséquilibre de classes** : une accuracy élevée peut masquer des erreurs sur la fraude.

12.2 Perspectives

— Fine-tuning d'un modèle NER sur un sous-ensemble annoté (schéma d'entités spécifique aux annonces).

— Modèle **hybride** : extraction + score de fraude + explications (indices textuels).

— Normalisation avancée (géocodage des lieux, standardisation des compétences via taxonomies).

13 Conclusion générale

Ce projet met en évidence l'intérêt de combiner méthodes classiques et modèles avancés pour la détection de fraudes textuelles. Les approches TF-IDF associées à la réduction de dimension et au KNN offrent un excellent compromis entre performance et simplicité. Les embeddings SBERT apportent une compréhension sémantique plus riche, au prix d'un coût computationnel plus élevé.

Les résultats montrent que le choix du modèle dépend du contexte :

pour des systèmes rapides et interprétables, les méthodes classiques sont très efficaces,

pour des cas plus complexes, les modèles pré-entraînés ouvrent des perspectives intéressantes.

Ce travail constitue une base solide pour des extensions futures, notamment des modèles hybrides ou des architectures neuronales spécialisées.

Références

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding*, 2019.

[2] N. Reimers, I. Gurevych. *Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks*, 2019.