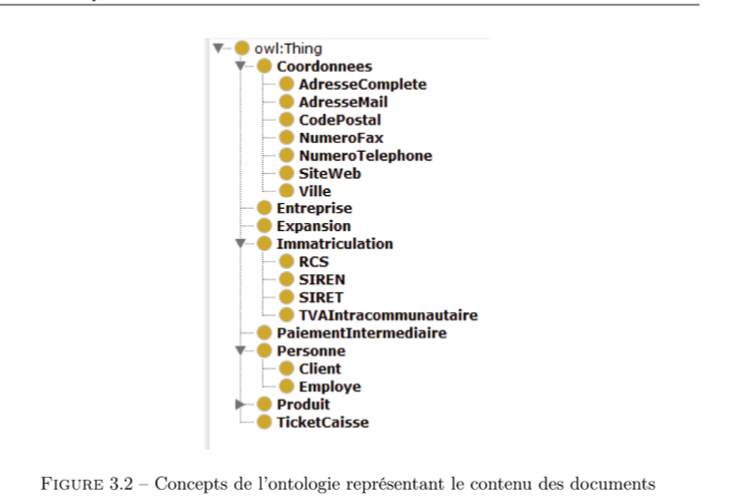
Lecture article ticket de caisse

Cette tache de détection de fraude au ticket de caisse peut se faire avec des techniques de traitement automatique du langage naturel (TALN), une approche basée sur la régression en tirant partit d’un modèle de langage pré-entrainé afin de représenter le contenu textuel, ils ont enrichi la représentation avec des entités et des relations basée sur une ontologie spécifique au domaine.

****

La contrefaçon est la production d’un document authentique par imitation et le faux est l’altération d’un ou plusieurs éléments sur un document authentique. L’un des principaux défis de la détection de la fraude de documents est le manque de données annotées librement disponibles.

La plupart des recherches en matière de fraude ce concentre principalement sur l’analyse d’image de document notamment grâce à la vision par ordinateur, effectivement l’image d’un document peut être modifiés sur le document numérique original ou dans sa version imprimée et numérisée tout cela a l’aide d’un logiciel d’édition d’image.

**Le modèle**

Le modèle de détection de fraude est pré-entrainé sur le modèle de CamemBERT qui est un modèle linguistique pré-entrainé de pointe pour le français lui-même basé sur le modèle RoBERTa.

Mettre des exemples de comment fonctionne CamemBERT.

**Les données**

Le dataset est composé de 998 images de ticket français et de leurs transcriptions associées, Ce jeu de données a été collecté pour fournir un corpus parallèle image/texte et un point de référence pour évaluer les méthodes de détection de fraude basées sur le texte. Les faux tickets sont le résultat d’ateliers de fraude, au cours desquels les participants ont reçu un ordinateur standard avec plusieurs logiciels d’édition d’images pour modifier manuellement les images et les transcriptions associées des tickets

**Situations réelles**

Des demandes de remboursement frauduleuses on pus être effectuées en modifiant le prix d’un article , son nom, le moyen de paiement, etc. La falsification peut également viser une extension indue de la garantie en modifiant la date.

D’autres falsifications peuvent impliquer l’entreprise émettrice dans le but de blanchir de l’argent.

**Pré-traitement et choix de l’entrée**

Ils ont expérimenté avec quatre types d’entrées :

**Texte :** le texte brut sans aucun pré-traitement

**Entités :** extraire les entités présentes sur la base d’une ontologie de ticket de caisse et les concatène avec un séparateur d’espace

**Texte + Entités :** enrichir le texte du ticket de caisse en introduisant des marqueurs spéciaux pour chaque type d’entité (tel que : entreprise, produit…) et remplace chaque entité dans le texte par son libellé entourer des marqueurs de son type d’entité.

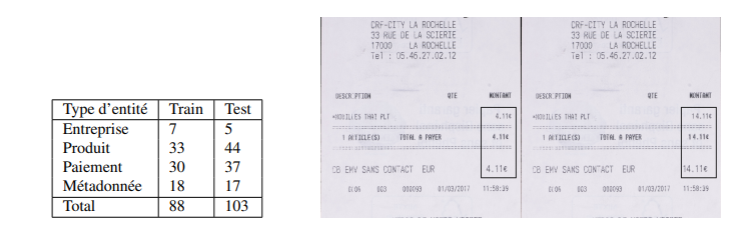
**Triplets :** basés sur la même ontologie, mais en extrayant également les relations sémantiques. (Relations sémantiques ?)

**Création de l’ontologie des tickets :**

Elle a été peuplée automatiquement avec des expressions régulières créées manuellement et basées sur les régularités des tickets de caisse. Par exemple, les produits et leurs prix ont été extraits des lignes du document terminées par le symbole « euro », ou l’utilisant comme séparateur décimal, à l’exclusion des lignes qui rapportent le total ou le paiement, pour ensuite crée automatiquement l’ontologie automatiquement grâce à la bibliothèque Owlready2.

**Extraction d’entités :**

Une entité extraite correspond à une instance d’une classe définie dans l’ontologie, qui définit sont type, ils ont annoté chaque entité fraudée, c’est-à-dire chaque entité ayant été modifier pendant la constitution du dataset (altération, ajout ou suppression), les modifications ne sont pas comptabilisées en elles-mêmes, seules les entités modifiées le sont : par exemple une date « 11/02/2023 » modifier en « 23/02/2024 » compte pour une entité modifiée, même si elle a subi deux modification graphique.



Le nombre d’entités modifiées est présenté dans le tableau de gauche et exemple de ticket de caisse modifié à droite.

La plupart des entités modifiées impliquent des montants d’argent (entités de produit et de paiement), même si ceux-ci ne sont pas toujours modifiés de manière cohérente, sur la figure de droite il y est présent trois fraudes le prix du produit, le prix total à payer et le montant payé qui ne comporte pas d’incohérence numérique.

**Expériences**

Premièrement il simule manuellement un vérificateur qui ne prend en compte que les incohérences simples telle qu’un écart entre le total et la somme des prix, entre le total et le total payé, ou entre la quantité, le prix unitaire et le prix du produit.

Ensuite il crée un classifieur de régression par machine à vecteur de support (SVM) avec des hyperparamètres par défaut comme notre modèle de base appliqué à la représentation fréquence des termes-fréquence inverse des documents (TF-IDF) des unigrammes et des bigrammes extraits des tickets en minuscules.

**Transformation d’un mot en TF-IDF :**

Les tickets de caisse sont analysés pour extraire des mots individuels (unigrammes) et des paires de mots (bigrammes), puis ces mots sont transformés en une forme spéciale appelée TF-IDF, qui aide à représenter leur importance dans les tickets. Cette représentation est effectuée en utilisant la fréquence des termes (TF) et la fréquence inverse du document (IDF).

Prenons un exemple simple pour illustrer la transformation d'un mot en TF-IDF.

Supposons que nous ayons un ensemble de documents, et nous voulons calculer le TF-IDF pour le mot "chat". Voici un exemple de quelques documents :

1. Document 1 : "Le chat est sur le tapis."
2. Document 2 : "Le chien et le chat jouent ensemble."
3. Document 3 : "Le chat dort paisiblement."

Maintenant, le calcul de TF-IDF se fait en deux étapes :

1. **TF (Term Frequency)** : C'est la fréquence du mot dans un document par rapport au nombre total de mots dans ce document.
   * Pour le document 1 : Le mot "chat" apparaît une fois sur un total de 5 mots. Donc, le TF pour "chat" dans le document 1 est 1/5.
   * Pour le document 2 : Le mot "chat" apparaît une fois sur un total de 6 mots. Donc, le TF pour "chat" dans le document 2 est 1/6.
   * Pour le document 3 : Le mot "chat" apparaît une fois sur un total de 4 mots. Donc, le TF pour "chat" dans le document 3 est 1/4.
2. **IDF (Inverse Document Frequency)** : C'est une mesure de l'importance du mot dans l'ensemble des documents. Cela permet de réduire le poids des mots très fréquents qui apparaissent dans de nombreux documents. Le calcul de IDF est effectué en prenant le logarithme de l'inverse de la fraction du nombre total de documents dans le corpus contenant le mot.
   * Le mot "chat" apparaît dans 3 documents sur un total de 3 documents. Donc, IDF pour "chat" est log(3/3) = log(1) = 0.

Maintenant, nous multiplions TF par IDF pour obtenir le score TF-IDF final pour le mot "chat" dans chaque document :

* Pour le document 1 : TF-IDF = (1/5) \* 0 = 0
* Pour le document 2 : TF-IDF = (1/6) \* 0 = 0
* Pour le document 3 : TF-IDF = (1/4) \* 0 = 0

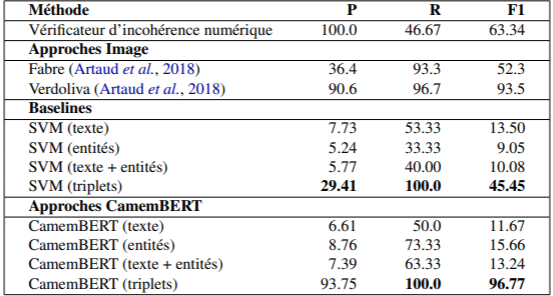
Dans cet exemple simplifié, le score TF-IDF pour le mot "chat" est 0 dans tous les documents, ce qui signifie que le mot "chat" n'est pas considéré comme très important dans cet ensemble de documents. Cependant, dans des ensembles de documents plus larges et plus variés, les scores TF-IDF peuvent varier pour refléter l'importance relative des mots dans chaque document.

Nous avons vu la manière mathématique d’aboutir à ce résultat mais il existe deux bibliothèques en python permettant ce dernier scikit-learn et NLTK (Natural Language Toolkit).

Il compare également les résultats à deux approches image déjà existantes, proposées dans la compétition Find it !. L’architecture Verdoliva est également basée sur un SVM et combine trois approche différentes : un module de détection de copier-coller, un module d’extraction (et de comparaison) de signature de caméra, et un module de détection de faux basée sur les caractéristiques locales de l’image, proposée a l’origine comme méthode de stéganalyse, et présente aussi les résultats de Fabre qui s’appuient sur un modèle pré-entrainé Resnet152 pour la classification.

**Résultat**

Le tableau ci-dessous présente les résultats de la classification binaire entres les classes Fraudé et Authentique. La classification étant très déséquilibrée, nous ne présentons que les résultats pour la classe Fraudé. Nous remarquons que les méthodes utilisant les Triplets comme entrée sont plus performantes que les autres, on pourra aussi remarquer que le recall est a 1 ce qui signifie que tous les tickets falsifiés sont retrouvés avec succès.



**Résumé**

- Le SVM est utilisé comme un type d'algorithme d'apprentissage automatique pour la classification des tickets de caisse en deux catégories : authentique ou frauduleux. Il est utilisé pour créer un classifieur basé sur les caractéristiques extraites des données des tickets.

- Le TALN est utilisé pour comprendre et extraire les informations textuelles des tickets de caisse. Dans ce cas, le TALN est utilisé pour traiter le texte brut des tickets de caisse, on peut donc déduire que le TALN travaillera sur les .txt de la BDD et que les données sur les qu’elle elle devra travailler et comprendre le sens sont sur les différents résultats des expressions régulières.

En combinant ces deux approches, le système peut utiliser à la fois les informations textuelles extraites des tickets de caisse et les caractéristiques des données pour détecter la fraude de manière efficace.

**Les critères de fraude à retenir :**

* Les incohérences simples telle qu’un écart entre le total et la somme des prix, entre le total et le total payé, ou entre la quantité, le prix unitaire et le prix du produit.
* Détection de copier-coller … suivre ce lien page 4 : chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://coria-taln-2023.sciencesconf.org/461873/document
* Modification du prix d’un article , son nom, le moyen de paiement. La falsification peut également viser une extension indue de la garantie en modifiant la date.
* Ce qui a été énumérée dans le résumé