# Détection de faux tickets de caisse à l'aide d'entités et de relations basées sur une ontologie de domaine

Beatriz Martínez Tornés[1] Emanuela Boros[1] Petra Gomez-Krämer[1]
Antoine Doucet[1] Jean-Marc Ogier[1]
(1) La Rochelle Université, L3i, F-17000, La Rochelle, France
{prénom.nom}@univ-lr.fr

RÉSUMÉ

Dans cet article, nous nous attaquons à la tâche de détection de fraude de documents. Nous considérons que cette tâche peut être abordée avec des techniques de traitement automatique du langage naturel (TALN). Nous utilisons une approche basée sur la régression, en tirant parti d'un modèle de langage pré-entraîné afin de représenter le contenu textuel, et en enrichissant la représentation avec des entités et des relations basées sur une ontologie spécifique au domaine. Nous émulons une approche basée sur les entités en comparant différents types d'entrée : texte brut, entités extraites et une reformulation du contenu du document basée sur des triplets. Pour notre configuration expérimentale, nous utilisons le seul ensemble de données librement disponible de faux tickets de caisse, et nous fournissons une analyse approfondie de nos résultats. Ils montrent des corrélations intéressantes entre les types de relations ontologiques, les types d'entités (produit, entreprise, etc.) et la performance d'un modèle de langage basé sur la régression qui pourrait aider à étudier le transfert d'apprentissage à partir de méthodes de traitement du langage naturel (TALN) pour améliorer la performance des systèmes de détection de fraude existants.

ABSTRACT

**Detecting Forged Receipts with Domain-specific Ontologies**

In this paper, we tackle the task of document fraud detection. We consider that this task can be addressed with natural language processing techniques. We treat it as a regression-based approach, by taking advantage of a pre-trained language model in order to represent the textual content, and by enriching the representation with domain-specific ontology-based entities and relations. We emulate an entity-based approach by comparing different types of input : raw text, extracted entities and a triple-based reformulation of the document content. For our experimental setup, we utilize the single freely available dataset of forged receipts, and we provide a deep analysis of our results in regard to the efficiency of our methods. Our findings show interesting correlations between the types of ontology relations, types of entities (product, company, etc.) and the performance of a regression-based language model that could help to study the transfer learning from natural language processing (NLP) methods to boost the performance of existing fraud detection systems.

MOTS-CLÉS : Détection de fraude de documents, modèle de langue.

KEYWORDS: Document fraud detection, language model.

# 1 Introduction

La falsification de documents est un problème très répandu, alors que la numérisation des documents permet un échange plus facile pour les entreprises et les administrations. Si l'on ajoute à cela la disponibilité de logiciels de traitement d'images et d'édition de documents ainsi que de scanners et d'imprimantes peu coûteux, les documents courent de nombreux risques d'être altérés ou contrefaits (Gomez-Kämer, 2021). La contrefaçon est la production d'un document authentique par imitation et le faux est l'altération d'un ou plusieurs éléments d'un document authentique. L'un des principaux défis de la détection de la fraude de documents est le manque de données annotées librement disponibles, car de nombreuses études autour de la fraude ne tiennent pas compte des documents réels et se concentrent sur les transactions (comme la fraude à la carte de crédit, la fraude à l'assurance ou même la fraude financière) (Behera & Panigrahi, 2015; Kowshalya & Nandhini, 2018; Rizki *et al.*, 2017).

La collecte de vrais faux documents est également difficile, car les vrais fraudeurs ne partageraient pas leur travail, et les entreprises ou administrations sont réticentes à révéler leurs failles de sécurité et ne peuvent pas partager des informations sensibles (Sidere *et al.*, 2017; Mishra & Ghorpade, 2018; Vidros *et al.*, 2017). De plus, le défi de travailler avec un corpus de documents administratifs potentiellement frauduleux est la rareté de la fraude ainsi que l'expertise humaine nécessaire pour repérer les documents frauduleux (Benchaji *et al.*, 2018; Li *et al.*, 2016; Carta *et al.*, 2019). S'intéresser aux documents réels réellement échangés par les entreprises ou les administrations est important pour que les méthodes de détection de fraude développées soient utilisables dans des contextes réels et pour que la cohérence des documents authentiques soit assurée. Cependant, ce type de document administratif contient des informations privées sensibles et n'est généralement pas mis à la disposition de la recherche (Behera & Panigrahi, 2015).

Ensuite, la plupart des recherches en matière de fraude de documents se concentrent sur l'analyse d'images de documents, car la plupart d'entre eux sont numérisés et échangés sous forme d'images par les entreprises et les administrations. La détection de falsification de documents est donc souvent définie comme une tâche de vision par ordinateur (Bertrand *et al.*, 2015; Cozzolino & Verdoliva, 2018; Fridrich & Kodovsky, 2012; Cozzolino *et al.*, 2014). L'image d'un document peut être modifiée de différentes manières à l'aide d'un logiciel d'édition d'images. La modification peut se faire dans le document numérique original ou dans la version imprimée et numérisée du document (Cruz *et al.*, 2018; James *et al.*, 2020). À cet égard, la compétition (Artaud *et al.*, 2018) a été, à notre connaissance, la seule tentative visant à encourager l'utilisation de méthodes de vision par ordinateur et de TALN pour la détection de documents falsifiés, en fournissant un corpus parallèle (image/texte) de tickets falsifiés librement accessibles. Toutefois, le nombre de participants était faible (cinq soumissions) et seule l'une d'entre elles intégrait des caractéristiques de contenu textuel sous la forme de modules de vérification basés sur des règles (l'examen des incohérences dans les prix des articles et le total à payer).

Nous considérons donc que le TALN pourrait être utilisé pour améliorer les performances de la détection des documents frauduleux en traitant les incohérences du faux lui-même (Artaud *et al.*, 2018). Ainsi, alors que les méthodes de vision par ordinateur s'appuient sur la recherche d'imperfections, soit en visant à détecter les irrégularités qui auraient pu se produire au cours du processus de modification (Bertrand *et al.*, 2013), soit en se concentrant sur l'identification de l'imprimante, afin de vérifier si le document a été imprimé par l'imprimante d'origine (Elkasrawi & Shafait, 2014; Mikkilineni *et al.*, 2005), les méthodes NLP pourraient combler le fossé entre les incohérences de l'image et du texte.

# 2 Notre approche

Nous basons notre modèle de détection des fraudes sur le modèle pré-entraîné CamemBERT (Martin *et al.*, 2020) qui est un modèle linguistique pré-entraîné de pointe pour le français basé sur le modèle RoBERTa (Liu *et al.*, 2019).

CamemBERT (Martin *et al.*, 2020) est une pile de couches Transformer (Vaswani *et al.*, 2017), où un bloc Transformer (encodeur) est une architecture d'apprentissage profond basée sur des mécanismes d'attention multitêtes avec des encastrements de position sinusoïdale. Comme indiqué précédemment, nous traitons la tâche de détection des fraudes comme une tâche de régression et un score numérique $s_x \in [0, 1]$ est donc attribué à l'exemple d'entrée $\{x_i\}_{i=1}^{l}$ pour quantifier son niveau de falsification, qui est défini comme $s_x = \sigma(f(\{x_i\})$ où $\sigma$ est la fonction sigmoïde $\sigma(z) = \frac{1}{1+e^{-z}}$ qui renvoie un score numérique $s_x \in [0, 1]$. Enfin, les valeurs prédites sont seuillées à 0,5.

## 2.1 Jeu de données de faux tickets de caisse

Le jeu de données librement disponible Find it ! [1] (Artaud *et al.*, 2017, 2018) que nous utilisons est composé de 998 images de tickets français et de leurs transcriptions associées. Ces transcriptions sont issues d'une reconnaissance optique de caractères et ont été corrigées manuellement de manière participative par les créateurs de Find it !. 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. Ainsi, le jeu de données contient des faux réalistes, correspondant à des situations réelles telles que des demandes de remboursement frauduleuses effectuées en modifiant le prix d'un article (Figure 1 (b)), 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. Les tickets ont été collectés localement dans le laboratoire de recherche où ils ont été développés, ce qui se traduit par une fréquence élevée de magasins à proximité. Bien que cela puisse être considéré comme un biais, nous estimons que cela reste proche d'un cas d'application réel, dans lequel une entreprise stocke les documents/factures qu'elle émet. Le jeu de données de 998 documents est divisé en 498 documents pour l'entraînement et 500 pour le test, chacun comportant 30 faux documents. Ainsi, les données sont déséquilibrées, selon une distribution réaliste. En effet, il y a typiquement moins de 5% de documents falsifiés dans les flux de documents, une distribution similaire aux valeurs aberrantes (Artaud *et al.*, 2018; Nadim *et al.*, 2019).

## 2.2 Prétraitements et choix de l'entrée

Afin de mieux explorer la nature spécifique semi-structurée des tickets, nous avons expérimenté avec quatre types d'entrées (présentées avec plus de détail dans la version longue de l'article (Martínez Tornés *et al.*, 2023)) :

1. **Texte** : le texte brut d'un ticket sans aucun prétraitement ;

1. http://findit.univ-lr.fr/download-the-dataset/

2. **Entités** : nous extrayons les entités présentes sur la base d'une ontologie de ticket de caisse et les concaténons avec un séparateur d'espace (par exemple« Carrefour ») comme décrit ci-dessous ;
3. **Texte + Entités** : nous enrichissons le texte du ticket en introduisant des marqueurs spéciaux pour chaque type d'entité (tel que, entreprise, produit, etc.) et remplaçons chaque entité dans le texte par son libellé entouré des marqueurs de son type d'entité (Boros *et al.*, 2021) ;
4. **Triplets** : basés sur la même ontologie, mais en extrayant également les relations sémantiques.

**L'ontologie** 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 « € », ou l'utilisant comme séparateur décimal, à l'exclusion des lignes qui rapportent le total ou le paiement. Le processus d'extraction a été exécuté comme une machine à états finis pour s'adapter à des structures plus variées, telles que des prix et des produits non alignés. L'ontologie a été peuplée dynamiquement à l'aide de la bibliothèque Python Owlready2 [2]. L'ontologie (Artaud, 2019) comporte des classes décrivant les informations présentes dans les tickets, telles qu'Entreprise, Adresse, Produit, etc. ainsi que des propriétés d'objet (relations entre ces classes) telles que a_adresse, contient. Nous notons que le ticket est une entité en soi, représentée par l'étiquette de son ID (une valeur numérique). L'ontologie définit également des propriétés de données, qui associent une instance à une valeur, comme *a_date*, *a_heure*, *a_montant_total*, *a_prix_total*, *a_nombre_darticles*, etc. [3]

**Extraction d'entités** Nous considérons qu'une entité extraite correspond à une instance d'une classe définie dans l'ontologie, qui définit son type. Nous avons annoté chaque entité fraudée, c'est-à-dire chaque entité ayant été modifiée pendant la constitution du dataset (altération, suppression ou ajout). Les modifications ne sont pas comptabilisées en elles-mêmes, seules les entités modifiées le sont : par exemple, une date « 11/02/2017 » altérée en « 10/02/2016 » compte pour une entité modifiée, même si elle a subi deux modifications graphiques. Le nombre d'entités modifiées est présenté dans le tableau 1 (a).

| Type d'entité | Train | Test |
|---|---|---|
| Entreprise | 7 | 5 |
| Produit | 33 | 44 |
| Paiement | 30 | 37 |
| Métadonnée | 18 | 17 |
| Total | 88 | 103 |

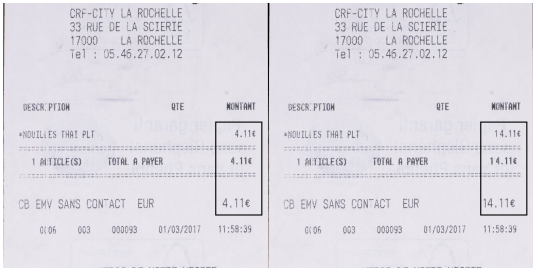


FIGURE 1 – **(a)** Distribution des entités modifiés. **(b)** Exemple de fraude sur le prix sur le ticket de droite.

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 : la Figure 1 (b) présente un

2. https://owlready2.readthedocs.io/en/v0.37/
3. Pour plus de détails et d'exemples, (Martínez Tornés *et al.*, 2023)

ticket comportant trois fraudes (le prix du produit, le total à payer et le montant payé) qui ne comporte pas d'incohérence numérique.

**Extraction des triplets** Afin d'aller au-delà des entités extraites et de fournir plus d'informations sur les relations entre les entités, nous avons choisi de les incorporer. Notre objectif était de mettre en évidence la structure sous-jacente des documents en énonçant explicitement les relations entre les entités. Nous avons veillé à supprimer les relations inverses, par exemple *a_fax* et *est_fax_de* en ne conservant qu'une seule de chaque paire. Nous avons également inclus les relations attributives, c'est-à-dire les propriétés des données, qui associent une entité à une valeur (numérique, date ou heure). Nous nous sommes appuyés sur ces triplets pour normaliser le contenu des tickets, en proposant une entrée (4) composée des triplets extraits en remplacement du texte extrait des tickets.

# 3 Expériences

Nous comparons notre modèle à deux méthodes *baseline*. Tout d'abord, nous considérons un *vérificateur d'incohérence numérique*, en simulant manuellement un vérificateur qui ne prend en compte que les incohérences numériques simples, sans s'appuyer sur des connaissances externes. Nous considérons comme une incohérence numérique simple tout 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. Deuxièmement, nous considérons 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.

Nous comparons également nos résultats à deux approches image existantes, proposées dans la compétition Find it ! (Artaud *et al.*, 2018). L'architecture Verdoliva (Cozzolino & Verdoliva, 2020, 2018) est également basée sur un SVM et combine trois approches différentes : un module de détection de fraude par copier-coller, basé sur Cozzolino *et al.* (2015), un module d'extraction (et de comparaison) de signatures de caméra (Cozzolino & Verdoliva, 2020, 2018), et un module de détection de faux basée sur les caractéristiques locales de l'image, proposée à l'origine comme méthode de stéganalyse (Cozzolino *et al.*, 2014). Nous présentons également les résultats proposés par Fabre (Artaud *et al.*, 2018), qui s'appuient sur un modèle pré-entraîné Resnet152 (He *et al.*, 2015) pour la classification.

**Résultats** Le tableau 1 détaille les résultats de la classification binaire entre 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, même dans leur représentation TF-IDF, le rappel est égal à un, ce qui signifie que tous les tickets falsifiés sont retrouvés avec succès.

Dans le cas des *Triplets*, nous n'avons observé que deux vrais tickets mal étiquetés. Pour l'un d'entre eux, le prix total est plutôt flou dans l'image, il a donc été corrigé manuellement dans la transcription avec « , » au lieu de « . » comme séparateur décimal. L'autre ticket authentique mal étiqueté présente un montant total élevé en comparaison avec le reste des tickets (plus de 87 euros). Ces irrégularités pourraient expliquer ces deux erreurs. En ce qui concerne la comparaison avec le

| Méthode | P | R | F1 |
|---|---|---|---|
| Vérificateur d'incohérence numérique | 100.0 | 46.67 | 63.34 |
| **Approches Image** | | | |
| Fabre (Artaud *et al.*, 2018) | 36.4 | 93.3 | 52.3 |
| Verdoliva (Artaud *et al.*, 2018) | 90.6 | 96.7 | 93.5 |
| **Baselines** | | | |
| SVM (texte) | 7.73 | 53.33 | 13.50 |
| SVM (entités) | 5.24 | 33.33 | 9.05 |
| SVM (texte + entités) | 5.77 | 40.00 | 10.08 |
| SVM (triplets) | **29.41** | **100.0** | **45.45** |
| **Approches CamemBERT** | | | |
| CamemBERT (texte) | 6.61 | 50.0 | 11.67 |
| CamemBERT (entités) | 8.76 | 73.33 | 15.66 |
| CamemBERT (texte + entités) | 7.39 | 63.33 | 13.24 |
| CamemBERT (triplets) | 93.75 | **100.0** | **96.77** |

TABLE 1 – Résultats de l'évaluation de la tâche de détection de tickets fraudés.

vérificateur d'incohérence numérique, nous avons constaté que notre approche est plus performante, comme en témoigne le rappel plus élevé. Cependant, il est important de noter la définition stricte de l'incohérence que nous avons utilisée : nous ne prenons en compte que les incohérences sur les valeurs numériques. La plupart des falsifications non détectées portent sur des valeurs modifiées de manière cohérente (le prix des articles est en accord avec le total et le montant payé, modification d'une date) et devraient être plus difficiles à repérer sans accès à des informations externes. Pourtant, dans l'ensemble de test, ces falsifications ne sont pas plausibles : par exemple, les tickets dans lesquels seule la date a été modifiée sont en fait attribués à une date impossible ou improbable, comme le 32/01, ou une année postérieure à l'arrêt de la collecte des données.

**Résultats par type de relation** Nous avons également analysé les résultats de la détection des faux tickets en utilisant un seul type de relation à la fois. Trois relations se sont distinguées par leurs résultats étonnamment parfaits (R=100) : *type*, *est_emis_par* et *a_paiement_total*. Ces résultats soulignent différents biais présents dans les données. Par exemple, près de 50 % des faux tickets ont été émis par Carrefour, qui ne représente que 30 % de l'ensemble des tickets. De plus, l'identifiant associé à chaque ticket de caisse n'est pas entièrement aléatoire, car les tickets de caisse sont au moins triés en fonction de l'entreprise qui les a émis. La relation attributive *a_paiement_total* permet d'obtenir des résultats très efficaces. Il faut s'attendre à un certain biais dans les valeurs numériques modifiées, comme la loi de Benford (Nigrini, 2012) qui décrit la distribution non normale des données numériques naturelles et qui a été utilisée dans la détection des fraudes comptables. Dans les nombres réels (tels que les prix, les nombres de population, etc.), le premier chiffre est susceptible d'être petit. En effet, les auteurs de l'ensemble de données signalent que l'utilisation de la loi de Benford pour rechercher des données numériques anormales permet d'obtenir un rappel de 70 % (Artaud, 2019). Ces résultats sont très encourageants quant à la capacité de notre approche à exploiter des informations statistiques, même sur des valeurs numériques, pour détecter les fraudes.

# 4 Conclusions

Cet article prouve que les méthodes basées sur le contenu sont capables de relever le défi de la détection de la fraude de documents au même niveau que les méthodes basées sur l'image. Notre objectif initial était d'établir une *baseline* et d'encourager les travaux futurs dans le domaine du NLP pour aborder la détection de la fraude de documents, et les résultats ont dépassé nos attentes. Notre approche basée sur le modèle pré-entraîné de CamemBERT considérant les relations entre les entités pour représenter le contenu des tickets atteint des valeurs de rappel élevées en exploitant efficacement les informations extraites des documents sous la forme de triplets.

# Remerciements

Ce travail a été soutenu par l'Agence de l'innovation de défense (AID), ainsi que le projet VERINDOC financé par la Région Nouvelle-Aquitaine.

# Références

ARTAUD C. (2019). *Détection des fraudes : de l'image à la sémantique du contenu. Application à la vérification des informations extraites d'un corpus de tickets de caisse*. PhD Thesis, University of La Rochelle.

ARTAUD C., DOUCET A., OGIER J.-M. & POULAIN D'ANDECY V. (2017). Receipt dataset for fraud detection. In *First International Workshop on Computational Document Forensics*.

ARTAUD C., SIDÈRE N., DOUCET A., OGIER J.-M. & POULAIN D'ANDECY V. (2018). Find it ! Fraud detection contest report. In *2018 24th International Conference on Pattern Recognition (ICPR)*, p. 13–18.

BEHERA T. K. & PANIGRAHI S. (2015). Credit card fraud detection : a hybrid approach using fuzzy clustering & neural network. In *2015 Second International Conference on Advances in Computing and Communication Engineering*.

BENCHAJI I., DOUZI S. & EL OUAHIDI B. (2018). Using genetic algorithm to improve classification of imbalanced datasets for credit card fraud detection. In *International Conference on Advanced Information Technology, Services and Systems*.

BERTRAND R., GOMEZ-KRÄMER P., TERRADES O. R., FRANCO P. & OGIER J.-M. (2013). A system based on intrinsic features for fraudulent document detection. In *2013 12th International Conference on Document Analysis and Recognition*, p. 106–110, Washington, DC, USA.

BERTRAND R., TERRADES O. R., GOMEZ-KRÄMER P., FRANCO P. & OGIER J.-M. (2015). A conditional random field model for font forgery detection. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, p. 576–580.

BOROS E., MORENO J. & DOUCET A. (2021). Event detection with entity markers. In *European Conference on Information Retrieval*, p. 233–240.

CARTA S., FENU G., RECUPERO D. R. & SAIA R. (2019). Fraud detection for e-commerce transactions by employing a prudential multiple consensus model. *Journal of Information Security and Applications*, **46**.

COZZOLINO D., GRAGNANIELLO D. & VERDOLIVA L. (2014). Image forgery detection through residual-based local descriptors and block-matching. In *2014 IEEE International Conference on Image Processing (ICIP)*.

COZZOLINO D., POGGI G. & VERDOLIVA L. (2015). Efficient dense-field copy–move forgery detection. *IEEE Transactions on Information Forensics and Security*, **10**(11).

COZZOLINO D. & VERDOLIVA L. (2018). Camera-based image forgery localization using convolutional neural networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*.

COZZOLINO D. & VERDOLIVA L. (2020). Noiseprint : A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, **15**, 144–159.

CRUZ F., SIDÈRE N., COUSTATY M., POULAIN D'ANDECY V. & OGIER J.-M. (2018). Categorization of document image tampering techniques and how to identify them. In *Pattern Recognition and Information Forensics - ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA, Revised Selected Papers*, p. 117–124.

ELKASRAWI S. & SHAFAIT F. (2014). Printer identification using supervised learning for document forgery detection. In *2014 11th IAPR International Workshop on Document Analysis Systems*, p. 146–150.

FRIDRICH J. & KODOVSKY J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, **7**(3).

GOMEZ-KÄMER P. (2021). Vérification de l'intégrité des documents. *Sécurité multimédia 2 : Biometrie, protection et chiffrement multimedia*, **2**, 71.

HE K., ZHANG X., REN S. & SUN J. (2015). Deep residual learning for image recognition. DOI : 10.48550/ARXIV.1512.03385.

JAMES H., GUPTA O. & RAVIV D. (2020). Ocr graph features for manipulation detection in documents.

KOWSHALYA G. & NANDHINI M. (2018). Predicting fraudulent claims in automobile insurance. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*.

LI Y., YAN C., LIU W. & LI M. (2016). Research and application of random forest model in mining automobile insurance fraud. In *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *ArXiv*, **abs/1907.11692**.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MARTÍNEZ TORNÉS B., BOROS E., GOMEZ-KRÄMER P., DOUCET A. & OGIER J.-M. (2023). Detecting Forged Receipts with Domain-specific Ontology-based Entities & Relations. In *Document Analysis and Recognition – ICDAR 2023*.

MIKKILINENI A. K., CHIANG P.-J., ALI G. N., CHIU G. T., ALLEBACH J. P. & DELP III E. J. (2005). Printer identification based on graylevel co-occurrence features for security and forensic applications. In *Security, Steganography, and Watermarking of Multimedia Contents VII*, volume 5681, p. 430–440 : International Society for Optics and Photonics.

MISHRA A. & GHORPADE C. (2018). Credit card fraud detection on the skewed data using various classification and ensemble techniques. In *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*.

NADIM A. H., SAYEM I. M., MUTSUDDY A. & CHOWDHURY M. S. (2019). Analysis of machine learning techniques for credit card fraud detection. In *2019 International Conference on Machine Learning and Data Engineering (iCMLDE)*, p. 42–47.

NIGRINI M. J. (2012). *Benford's Law : Applications for forensic accounting, auditing, and fraud detection*, volume 586. John Wiley & Sons.

RIZKI A. A., SURJANDARI I. & WAYASTI R. A. (2017). Data mining application to detect financial fraud in indonesia's public companies. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*.

SIDERE N., CRUZ F., COUSTATY M. & OGIER J.-M. (2017). A dataset for forgery detection and spotting in document images. In *2017 Seventh International Conference on Emerging Security Technologies (EST)*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

VIDROS S., KOLIAS C., KAMBOURAKIS G. & AKOGLU L. (2017). Automatic detection of online recruitment frauds : Characteristics, methods, and a public dataset. *Future Internet*, **9**(1).