Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

Jeu de données de tickets de caisse pour la détection de fraude documentaire

Beatriz Martínez Tornés[1] Théo Taburet[1] Emanuela Boros[1] Kais Rouis[1] Petra Gomez-Krämer[1] Nicolas Sidère[1] Antoine Doucet[1] Vincent Poulain d'Andecy[2]
(1) La Rochelle Université, L3i, F-17000, La Rochelle, France
{prénom.nom}@univ-lr.fr
(2) Yooz, 1 Rue Fleming, 17000 La Rochelle, France
Vincent.PoulaindAndecy@getyooz.com

RÉSUMÉ

L'utilisation généralisée de documents numériques non sécurisés par les entreprises et les administrations comme pièces justificatives les rend vulnérables à la falsification. En outre, les logiciels de retouche d'images et les possibilités qu'ils offrent compliquent les tâches de la détection de fraude d'images numériques. Néanmoins, la recherche dans ce domaine se heurte au manque de données réalistes accessibles au public. Dans cet article, nous proposons un nouveau jeu de données pour la détection des faux tickets contenant 988 images numérisées de tickets et leurs transcriptions, provenant du jeu de données SROIE (scanned receipts OCR and information extraction). 163 images et leurs transcriptions ont subi des modifications frauduleuses réalistes et ont été annotées. Nous décrivons en détail le jeu de données, les falsifications et leurs annotations et fournissons deux *baselines* (basées sur l'image et le texte) sur la tâche de détection de la fraude.

ABSTRACT

Receipt Dataset for Document Forgery Detection

The widespread use of unsecured digital documents by companies and administrations as supporting documents makes them vulnerable to forgeries. Moreover, image editing software and the capabilities they offer complicate the tasks of digital image forensics. Nevertheless, research in this field struggles with the lack of publicly available realistic data. In this paper, we propose a new receipt forgery detection dataset containing 988 scanned images of receipts and their transcriptions, originating from the scanned receipts OCR and information extraction (SROIE) dataset. 163 images and their transcriptions have undergone realistic fraudulent modifications and have been annotated. We describe in detail the dataset, the forgeries and their annotations and provide two baselines (image and text-based) on the fraud detection task.

MOTS-CLÉS : Fraude documentaire, jeu de données, détection de fraude.

KEYWORDS: Document forgery, dataset, fraud detection.

1 Introduction

La détection automatique de fraudes est devenue une tâche inévitable dans les flux de documents des entreprises, car l'acceptation de documents falsifiés peut servir comme support à d'autres types de fraude. Par exemple, un fraudeur peut grâce à des faux documents s'assurer une usurpation d'identité

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 4 : articles déjà soumis ou acceptés en conférence internationale, pages 140–147.

ou à l'obtention d'un prêt pour financer des activités criminelles telles que des attaques terroristes. Cependant, les travaux de recherche proposés manquent de généralité, car ils sont très spécifiques à un certain type ou méthode de falsification, et il en va de même pour les ensembles de données disponibles. L'un des principaux défis de la détection de la fraude documentaire est le manque de données annotées librement disponibles. En effet, la collecte de documents frauduleux est entravée par la réticence des fraudeurs à partager leur travail, comme on peut s'y attendre dans toute activité illégale, ainsi que par les contraintes qui pèsent sur les entreprises et les administrations pour partager des données sensibles (Sidere *et al.*, 2017; Mishra & Ghorpade, 2018; Vidros *et al.*, 2017). En outre, de nombreuses études sur la fraude ne se concentrent pas sur les documents eux-mêmes, mais sur les transactions, telles que la fraude à l'assurance, la fraude à la carte de crédit ou la fraude financière (Behera & Panigrahi, 2015; Kowshalya & Nandhini, 2018; Rizki *et al.*, 2017). Nous tentons donc de combler ce fossé entre le manque d'ensembles de données de détection de falsifications disponibles publiquement et l'absence de contenu textuel exploitable, en construisant un nouvel ensemble de données génériques pour la détection de falsifications basé sur des images de documents réels. Nous avons basé l'ensemble de données sur jeu de données existant de tickets scannés (SROIE) qui a été initialement proposé pour des tâches d'extraction d'informations et qui contient des images et du texte. Nous avons altéré les images en utilisant plusieurs méthodes d'altération (copier-coller, imitation de texte, suppression d'informations et modification de pixels) et modifié les transcriptions en conséquence. Nous fournissons les images ainsi que les transcriptions permettant une analyse en texte seul.

2 Construction d'un jeu de données pour la détection de faux tickets

S'intéresser aux documents réellement échangés par les entreprises ou les administrations est essentiel pour que les méthodes de détection de fraude développées soient utilisables dans des contextes réels et que la cohérence des documents authentiques soit assurée. Cependant, ces documents administratifs contiennent des informations privées sensibles et ne sont généralement pas mis à disposition de la recherche (Artaud *et al.*, 2018). Nous considérons la tâche de détection de la fraude sur les tickets, car les tickets ne contiennent pas d'informations sensibles et ont une structure très similaire à celle des factures. Ainsi, des scénarios réalistes peuvent être associés à la falsification de tickets, tels que le remboursement de frais de voyage (gagner un peu d'argent supplémentaire, produits non remboursés), et la preuve d'achat (pour l'assurance, pour la garantie).

2.1 Données du SROIE

L'ensemble de données a été choisi comme point de départ pour créer l'ensemble de données de falsification. Il a été créé à l'origine pour l'OCR de tickets numérisés et l'extraction d'informations (SROIE) dans le cadre d'une compétition ICDAR 2019 et contient 1 000 images de tickets numérisés accompagnées de leurs transcriptions, issues de la vérité terrain de la compétition.

Une caractéristique des tickets scannés originaux du SROIE est que certains ont été modifiés, soit numériquement, soit manuellement. Ces modifications ne sont pas considérées comme des faux. Même si les documents ont été modifiés, ils restent authentiques, car ils n'ont pas été falsifiés. Ces annotations conviennent à notre étude de cas, car la plupart d'entre elles sont des notes spécifiques au

contexte que l'on trouve dans les applications de documents réels. Par exemple, certaines annotations correspondent à des notes laissées sur les tickets, telles que « staff outing » pour décrire la nature de l'événement (figure 2), des chiffres qui peuvent décrire une mission ou un numéro de dossier (figures 1 et 4), des noms ou des marques pour mettre en évidence des informations clés sur le document, telles que le prix dans la figure 3. Nombre de ces annotations peuvent même provenir du processus de collecte de l'ensemble de données et sont difficiles à interpréter sans plus d'informations contextuelles (noms, numéros, etc.).

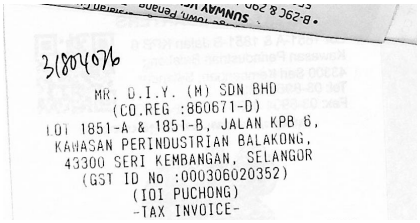

FIGURE 1 – Insertion manuelle de chiffres.

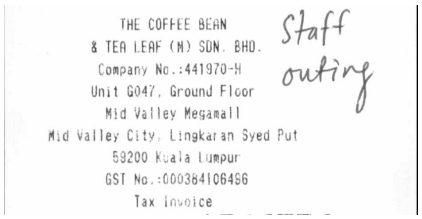

FIGURE 2 – Note sur le ticket.

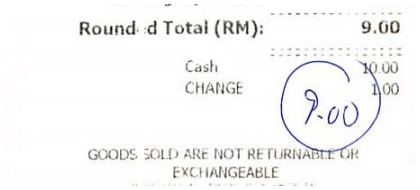

FIGURE 3 – Mise en valeur du total.

FIGURE 4 – Insertion numérique de chiffres.

Ces modifications apportées à des documents authentiques posent le problème difficile de la détection de la fraude, qui consiste à faire la distinction entre une modification frauduleuse et une modification non malveillante. Une modification frauduleuse se caractérise non seulement par la mauvaise intention de son auteur, mais aussi par le fait qu'elle modifie des caractéristiques structurelles ou significatives cruciales du document qui peuvent être utilisées pour déformer le sens du document original.

Afin d'évaluer l'impact de ces modifications, nous avons annoté manuellement les tickets en fonction du type de modifications qu'ils ont subies. Nous définissons une annotation numérique comme un cas particulier récurrent d'une séquence de numéros ou de noms ajoutés numériquement dans plusieurs en-têtes de tickets, comme dans la figure 4. Nous considérons qu'il y a eu une annotation manuelle (figures 1, 2, et 3) s'il y a une note manuscrite de quelque type que ce soit sur le ticket (mots, coches, zones surlignées ou soulignées, etc.), ainsi que des tampons. Nous avons remarqué que les annotations numériques sont reportées dans les transcriptions, alors que les annotations manuelles ne le sont pas. Au total, nous avons compté 54 tickets contenant une annotation numérique, 500 comptant une annotation manuelle et 34 contenant une annotation numérique et une annotation manuelle.

2.2 Campagne de fraude

Afin de fournir des faux aussi réalistes que possible, nous avons organisé plusieurs ateliers de falsification similaires à ceux réalisés par les ensembles de données falsifiées pseudo-réalistes (Artaud *et al.*, 2018; Sidere *et al.*, 2017). Les 19 participants étaient des volontaires, principalement issus du

monde de l'informatique, même si nous avons tenté d'élargir la portée de notre projet à différents niveaux de compétence et d'expertise en matière de documents numériques et de logiciels de retouche d'images. L'objectif n'était pas de créer un ensemble de données d'experts en falsification, mais d'avoir une représentation réaliste de différentes compétences et temps consacré. Les participants n'ont pas reçu de directives spécifiques sur les outils et les techniques à utiliser, afin qu'ils puissent utiliser ce avec quoi ils étaient le plus à l'aise. Cinq logiciels différents ont été utilisés : `aperçu` (15 documents), `paint` (70), `paint3d` (10), `GIMP` (65) et `kolourpaint` (3).

Modification du texte et de l'image Les participants ont reçu des exemples et des scénarios pour commencer, tels que le remboursement de frais de mission, la preuve d'achat pour l'assurance ou pour la garantie (par exemple, date trop ancienne). Il leur a été demandé de modifier l'image ainsi que son fichier texte associé (transcription).

Annotation des fraudes Ensuite, les participants ont été invités à annoter les fraudes qu'ils venaient de réaliser à l'aide de l'annotateur d'images VGG [1]. Ces annotations sont fournies avec l'ensemble de données au format JSON, comme le montre l'exemple suivant :


```
{'filename': 'X51005230616.png', 'size': 835401, 'regions':
[{'shape_attributes': {'name': 'rect', 'x': 27, 'y': 875, 'width': 29,
'height': 43}, 'region_attributes': {'Modified area': {'IMI': True},
'Entity type': 'Product', 'Original area': 'no'}},
{'shape_attributes': {'name': 'rect', 'x': 458, 'y': 883, 'width': 35,
'height': 37},'region_attributes': {'Modified area': {'IMI': True},
'Entity type': 'Product', 'Original area': 'no'}}],
'file_attributes': {'Software used': 'paint', 'Comment': ''}}
```


Le processus a consisté, d'une part, à localiser les zones modifiées en définissant les régions rectangulaires concernées et, d'autre part, à décrire le type de falsification selon la nomenclature proposée dans (Cruz *et al.*, 2019) (copier-coller à partir du même document, copier-coller à partir d'un autre document, suppression d'informations et imitation). Nous incluons un type de falsification supplémentaire (PIX) pour toutes les modifications « à main levée » (James *et al.*, 2020). Nous avons donc proposé les types de falsification suivants :

- **CPI** : Copier et coller à l'intérieur du document, c'est-à-dire copier une partie de l'image (un caractère, un mot entier, une séquence de mots, etc.) et la coller dans la même image ;
- **CPO** : Copier-coller en dehors du document, c'est-à-dire copier une partie de l'image (un caractère, un mot entier, une suite de mots, etc.) et la coller dans un autre document ;
- **IMI** : Zone de texte imitant la police, utilisant un outil d'insertion de texte pour remplacer ou ajouter un texte ;
- **CUT** : Supprimer un ou plusieurs caractères, sans les remplacer ;
- **PIX** : Modification par pixel, pour toutes les modifications effectuées « à main levée » avec un outil de type pinceau pour introduire une modification (par exemple, transformer un caractère en un autre en ajoutant une ligne) ;
- ***Autre*** : Utilisation de filtres ou d'autres éléments (à préciser dans les commentaires).

Annotation des entités modifiées Les participants ont également été invités à identifier le type d'entité altérée pour chaque zone modifiée de la liste suivante :

1. https://www.robots.ox.ac.uk/$\sim$vgg/software/via/

— **Entreprise** : Informations relatives à l'entreprise (adresse, téléphone, nom) ;
— **Produit** : Informations relatives à un produit (nom, prix, suppression ou ajout d'un produit) ;
— **Total/Paiement** : Prix total, mode de paiement ou montant payé ;
— **Metadonnées** : Date, heure.

Post-traitement Toutes les annotations fournies pour les faux tickets sont manuelles. Afin de nous assurer que les zones modifiées ont été correctement annotées par les participants, nous avons corrigé manuellement toutes les annotations. Ces corrections ont été effectuées en comparant les documents falsifiés aux originaux.

2.3 Description du dataset

Le jeu de données contient 988 images PNG avec leurs transcriptions correspondantes dans un format texte. Les données peuvent être téléchargées à l'adresse `http://l3i-share.univ-lr.fr/2023Finditagain/findit2.zip`. Nous proposons une répartition des données entre les ensembles d'entraînement, de validation et de test afin de permettre une comparaison entre les différentes méthodes. La répartition des données est décrite dans le tableau 1, avec les décomptes des faux effectués pendant la campagne de falsification (Section 2.2) ainsi que les annotations présentes dans les tickets authentiques (Section 2.1).

	Entraînement	**Validation**	**Test**	**Total**
Nombre de tickets	577	193	218	988
Nombre de tickets fraudés	94	34	35	163
Nombre de tickets avec une annotation numérique	34	9	11	54
Nombre de tickets avec une annotation manuelle	305	86	109	500

TABLE 1 – Répartition des données.

Au total, 455 zones différents ont été modifiés dans 163 documents. Le tableau 2 détaille le nombre de modifications effectuées par type : une même zone peut avoir été affectée par plus d'un type de modification. En ce qui concerne les entités, la plupart des modifications ont porté sur les informations relatives au total ou au paiement. Le tableau montre également que la technique de falsification la plus utilisée est le CPI.

Type de fraude	**Décompte**	**Type d'entité**	**Décompte**
CPI	353	Total/paiement	234
IMI	36	Produit	95
CUT	36	Métadonnées	82
PIX	33	Entreprise	26
CPO	10	Autre	18

TABLE 2 – Description des zones modifiées.

3 Baselines proposées

Bag-of-words (BoW) & régression logistique Tout d'abord, nous avons choisi ce modèle de classification sur une représentation en sac-de-mots, car il sert généralement de modèle de base et

peut-être utilisé comme référence pour évaluer les résultats et avoir un premier aperçu de la difficulté de la tâche. Nous considérons le modèle le plus couramment utilisé pour une *baseline* simple et rapide : la régression logistique (RL).

ChatGPT Le modèle récent créé par OpenAI [2], proposé en novembre 2022, a suscité une grande attention dans les communautés universitaires et industrielles, et a été rapidement adopté par tout types d'utilisateurs, non seulement en raison de son impressionnante capacité à engager des conversations, mais aussi de sa capacité à répondre aux questions de suivi, à paraphraser, à corriger les fausses déclarations et à décliner les demandes inappropriées (Guo *et al.*, 2023). Nous étions donc curieux de comparer les réponses d'un expert humain et de ChatGPT à la même question (Askell *et al.*, 2021). Nous avons utilisé l'invite suivante [3] :

```
Extract the locations (LOC), products (PROD) and prices (PRI)
from the following receipt and tell me if it's fraudulent:{receipt}
```

Comme les réponses dans le texte libre ne correspondent pas à des résultats de classification binaire, nous les alignons selon deux configurations :

- **Strict** : Seules les réponses exprimant des doutes précis ou des éléments notables concernant le ticket ou sa légitimité ont été classées comme fausses. Pour sept reçus seulement, la réponse indiquait explicitement qu'un élément était « digne d'intérêt », « suspect » ou semblait « frauduleux ».
- **Relâché** : Toutes les réponses qui ne se prononçaient pas explicitement en faveur de la classe authentique on été considérées comme des faux tickets.

Les résultats [4] sont présentés dans la table 3. Le jeu de données étant très déséquilibré, nous ne présentons que les résultats de classification binaire pour la classe « Faux ». L'approche de classification de texte et son très faible rappel montrent à quel point elle est insuffisante : seuls quatre faux tickets ont été correctement étiquetés par le classificateur de texte.

Méthode	Précision	Rappel	F1-score
Classification de texte (BoW + LR)	40.00	11.43	17.78
ChatGPT (strict)	14.69	**88.57**	25.20
ChatGPT (relâché)	18.33	62.86	**28.39**

TABLE 3 – Résultats.

4 Conclusions et perspectives

Cet article présente le jeu de données de tickets librement disponible [5] pour la détection de documents fraudés, contenant à la fois des images et leur transcription de 988 tickets. Il fournit également des annotations sémantiques sur les zones modifiées, ainsi que des détails sur les techniques de falsification utilisées et leurs zones de délimitation. L'ensemble de données peut donc être utilisé pour des tâches de classification et de localisation. Nous pensons que cet ensemble de données peut

2. https://openai.com/blog/chatgpt/
3. L'invite et les réponses ont été traduites de l'anglais par nous.
4. Pour plus de détails et d'expériences sur les images des documents, cet article est une traduction d'un article accepté en version longue (Martínez Tornés *et al.*, 2023).
5. http://l3i-share.univ-lr.fr/2023Finditagain/findit2.zip

constituer une ressource intéressante pour la communauté de détection des faux documents. Les expériences présentées peuvent être considérées comme un point de départ pour comparer avec d'autres méthodes, en particulier le développement d'approches TALN pour l'authentification de documents, ainsi que des approches multimodales. En effet, si l'analyse des documents et la détection de fraude sont majoritairement abordés comme des problématiques de vision par ordinateur, la prise en compte du contenu ainsi que sa cohérence et sa plausibilité est une perspective qui nous semble prometteuse, mais qui est longtemps restée limitée par le manque de données réalistes (ou pseudo-réalistes).

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. Nous tenons également à remercier les participants à la campagne de fraude pour leur contribution.

Références

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.

ASKELL A., BAI Y., CHEN A., DRAIN D., GANGULI D., HENIGHAN T., JONES A., JOSEPH N., MANN B., DASSARMA N. *et al.* (2021). A general language assistant as a laboratory for alignment. *arXiv preprint arXiv :2112.00861*.

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*.

CRUZ F., SIDÈRE N., COUSTATY M., POULAIN D'ANDECY V. & OGIER J.-M. (2019). Categorization of document image tampering techniques and how to identify them. In *International Conference on Pattern Recognition*, p. 117–124 : Springer.

GUO B., ZHANG X., WANG Z., JIANG M., NIE J., DING Y., YUE J. & WU Y. (2023). How close is chatgpt to human experts ? comparison corpus, evaluation, and detection. *arXiv preprint arXiv :2301.07597*.

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)*.

MARTÍNEZ TORNÉS B., TABURET T., ROUIS K., BOROS E., GOMEZ-KRÄMER P., SIDERE N., DOUCET A. & POULAIN D'ANDECY V. (2023). Receipt Dataset for Document Forgery Detection. In *2023 International Conference on Document Analysis and Recognition (ICDAR)*.

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)*.

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)*.

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).