Detecting forged receipts with domain-specific ontology-based entities & relations

Beatriz Martínez Tornés, Emanuela Boros, Petra Gomez-Krämer, Antoine Doucet, Jean-Marc Ogier

▶ **To cite this version:**

Beatriz Martínez Tornés, Emanuela Boros, Petra Gomez-Krämer, Antoine Doucet, Jean-Marc Ogier. Detecting forged receipts with domain-specific ontology-based entities & relations. ICDAR, Aug 2023, San José, United States. pp.184-199, 10.1007/978-3-031-41682-8_12 . hal-04296021

HAL Id: hal-04296021
https://hal.science/hal-04296021v1

Submitted on 20 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Detecting Forged Receipts with Domain-specific Ontology-based Entities & Relations

Beatriz Martínez Tornés[1 [0000−0002−7820−640X]],
Emanuela Boros[1 [0000−0001−6299−9452]], Antoine Doucet[1 [0000−0001−6160−3356]],
Petra Gomez-Krämer[1 [0000−0002−5515−7828]], and
Jean-Marc Ogier[1 [0000−0002−5666−475X]]

University of La Rochelle, L3i, F-17000, La Rochelle, France
firstname.lastname@univ-lr.fr

Abstract. 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 (e.g., has_address, amounts_to), 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.

Keywords: Fraud detection · Language models · Ontology.

1 Introduction

Document forgery is a widespread problem, while document digitization allows for easier exchange for companies and administrations. Coupled with the availability of image processing and document editing software as well as cost-effective scanners and printers, documents face many risks to be tampered with or counterfeited [21], where counterfeiting is the production of a genuine document from scratch by imitation and forgery is the alteration (tampering) of one or more elements of an authentic document.

First, one of the main challenges of document fraud detection is the lack of freely available annotated data, as many studies around fraud do not consider the actual documents and focus on the transactions (such as credit card fraud, insurance fraud, or even financial fraud) [6, 27, 39]. Collecting real forged documents is also difficult because real fraudsters would not share their work,

and companies or administrations are reluctant to reveal their security breaches and cannot share sensitive information [42, 34, 46]. Moreover, the challenge of working with a corpus of potentially fraudulent administrative documents is the scarcity of fraud as well as the human expertise required to spot the fraudulent documents [7, 30, 12]. Taking an interest in real documents actually exchanged by companies or administrations is important for the fraud detection methods developed to be usable in real contexts and for the consistency of authentic documents to be ensured. However, this type of administrative document contains sensitive private information and is usually not made available for research [6].

Second, most of the research in document forensics is focused on the analysis of images of documents, as most of these are scanned and exchanged as images by companies and administrations. Document forgery detection is thus often defined as a tampering detection computer vision (CV) task [9, 15, 20, 13]. A document image can be tampered with in different ways with the help of image editing software. The modification can be done in the original digital document or in the printed and digitized version of the document, which is usually a scanned document, as the mobile-captured document contains too many distortions. Thus, the document can then be printed and digitized again to hide the traces of the fraud [18, 24].

In these regards, the *Find it!* competition [4] was, to the best of our knowledge, the only attempt to encourage both CV and natural language processing (NLP) methods to be used for document forgery detection, by providing a freely-available parallel (image/text) forged receipt corpora. However, the number of participants was low (five submissions) and only one of them incorporated content features in the form of rule-based check modules (i.e., looking at inconsistencies in article prices and the total to pay), which proved to be rather effective (an F1 of 0.638).

We, thus, consider that NLP and knowledge engineering (KE) could be used to improve the performance of fraudulent document detection by addressing the inconsistencies of the forgery itself [4]. Hence, while CV methods rely on finding imperfections, by either aiming to detect irregularities that might have occurred during the modification process [8] or by focusing on printer identification, in order to verify if the document has been printed by the original printer [19, 33], NLP methods could bridge the gap between image and textual inconsistencies [43]. We experiment with a pre-trained language model regression-based approach while also tailoring the textual input by generating ontology-based entities and relations in documents in order to provide more semantic content of a forged French receipt dataset [3, 4]. Our findings show interesting correlations between the types of ontology relations (e.g., has_contactDetail, has_address), the types of entities (product, company, etc.) and the performance of our approach that could foster further research and help to study the transfer learning from NLP methods to boost the performance of existing fraud detection systems.

The paper is organized as follows. Section 2 presents the state of the art with CV-based fraud detection methods, as well as in NLP. Section 3 introduces the forgery detection receipt dataset we used in this study. Our semantic-aware

approach is described in Section 4, focusing on our alternative textual ontology-based inputs. We then present the experiments and results in Section 5. Finally, Section 6 states the conclusions and future work.

2 Related Work

Computer Vision-based Fraud Detection Most of the research in document forensics is part of the field of computer vision (CV) [9, 15, 20, 13]. As a result, most fraud detection datasets are not focused on the semantic content and its alteration. Therefore, they are not usable by a textual approach. Indeed, most datasets are synthetically curated to evaluate a particular CV approach [34, 6, 39]. For a concrete example, the automatically generated documents in [8] were also automatically tampered with (in terms of noise and change in size, inclination, or position of some characters). The payslip corpus was created by randomly completing the various fields required for this type of document [42]. Datasets used for source scanner (or printer) identification consist of the same documents scanned (or printed) by different machines, without any actual content modification [38]. These datasets are suitable for image-based approaches [14, 16, 17], but are not relevant for content analysis approaches, as the forged documents are as inconsistent as the authentic ones. Some works focus on the detection of graphical indices of the document modification such as slope, size, and alignment variations of a character with respect to the others [8], font or spacing variations of characters or in a word [9], the variation of geometric distortions of characters introduced by the printer [41], the text-line rotation and alignment [44] or an analysis of the document texture [17]. The authors of [2] use distortions in the varying parts of the documents (not the template ones) through pair-wise document alignment to detect forgery. Hence, the methods need several samples of a class (template). A block-based method for copy and move forgery detection was also proposed which is based on the detection of similar characters using Hu and Zernike moments, as well as PCA and kernel PCA combined with a background analysis [1]. The principle of detecting characters is similar to that of [8] using Hu moments. The method of [17] is the more generic, as it is not related to a certain type of tampering. It is based on an analysis of LBP textures to detect discontinuities in the background and residuals of the image tampering. Due to the difficulty of the task and the lack of generality of these methods, only a few works have been proposed for this task.

Natural Language Processing-based Fraud Detection Since most of the research in document forensics is focused on the analysis of images of documents, NLP-based fraud detection suffers from a lack of previous work. However, existing fraud detection approaches, in a broader sense than document forgery, mainly focus on supervised machine learning (e.g., neural networks, bagging ensemble methods, support vector machine, and random forests) based on manual feature engineering [34, 6, 27, 30, 25]. Knowledge graph embeddings-based approaches have also been proposed to tackle content-based fraud detection in these types of documents [**?**,40]. However, this approach congregated all the documents in order

to learn a representation of the different extracted semantic relations and used graph-based methods in order to add data from external sources. Moreover, the approach did not prove to be efficient, compared to CV state-of-the-art results. Recently, language models based on BERT [28, 22] have been developed and proven to outperform state-of-the-art results in anomaly detection in system logs, and records of events generated by computer systems. However, the methods are sequential, and cannot be applied in receipt fraud detection where segments of text can be erased (e.g., the removal of a purchased product and its price).

3 Forged Receipt Dataset

The freely available dataset [3, 4] that we utilize is composed of 998 images of French receipts and their associated optical character recognition (OCR) results. It was collected to provide an image/text parallel corpus and a benchmark to evaluate our text-based methods for fraud detection. The forged receipts are the result of tampering workshops, in which participants were given a standard computer with several image editing software to manually alter both images and associated OCR results of the receipts. Thus, the dataset contains realistic forgeries, consistent with real-world situations such as fraudulent refund claims made by modifying the price of an item as shown in Figure 1 (a), its label, the means of payment, etc. The forgery can also target an undue extension of warranty by modifying the date (however, unless the date is implausible, as, in the example shown in Figure 2 (b), there is no semantic inconsistency). Other forgeries can involve the issuing company with the aim of money laundering, as in the example in Figure 3 (c) which produces a false invoice to a false company.

Fig. 1. Price forgery.

Fig. 2. Date forgery.

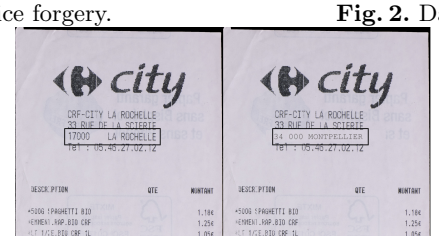


Fig. 3. Address forgery.

The receipts were collected locally in the research laboratory they were developed, which results in a high frequency of stores in the vicinity. Although this can be seen as a bias, we consider it remains close to a real application case, in which a company stores the documents/invoices it emits. Given the quality of most receipts, in terms of ink, paper and care to avoid crumpling, the automated OCR results were not usable. They were thus manually corrected, both automatically (to tackle recurring errors such as "€" symbols at the end of lines) and manually. The manual correction was performed participatively[1]. The dataset of 998 documents is split into 498 documents for training and 500 for testing, each with 30 forged documents. Thus, the data is imbalanced, according to a realistic distribution of the data. Indeed, there is typically less than 5% of forged documents in document flows, a distribution similar to outliers [4, 36].

4 Language Model Regression-based Approach

We base our fraud detection model on the pre-trained model CamemBERT [32] which is a state-of-the-art pre-trained language model for French based on the RoBERTa model [31].

4.1 Model Description

CamemBERT [32] is a stack of Transformer [45] layers, where a Transformer block (encoder) is a deep learning architecture based on multi-head attention mechanisms with sinusoidal position embeddings. In detail, let $\{x_i\}_{i=1}^{l}$ be a token input sequence consisting of l words, denoted as $\{x_i\}_{i=1}^{l} = \{x_1, x_2, \dots x_i, \dots x_l\}$, where $x_i (1 \leq x_i \leq l)$ refers to the i-th token in the sequence of length l. CamemBERT, similarly to other language models, expect the input data in a specific format: a special token, [SEP], to mark the end of a sentence or the separation between two sentences, and [CLS], at the beginning of a text, used for classification or regression tasks. We chose CamemBERT's [CLS] token output vector $[CLS]$, denoted by $CamemBERT_{[CLS]}$, as the input of the model and then, apply $CamemBERT$ for further fine-tuning: $f(\{x_i\}_{i=1}^{l}) = CamemBERT_{[CLS]}W_t$ where $W_t \in R^{d_{model} \times 1}$ are the learnable parameters of the linear projection layer and d_{model} is the hidden state dimension of CamemBERT.

As previously mentioned, we treat the fraud detection task as a regression task and thus a numeric score $s_x \in [0, 1]$ is assigned to the input example $\{x_i\}_{i=1}^{l}$ for quantifying its forging level, which is defined as $s_x = \sigma(f(\{x_i\})$ where σ is the sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$ that returns a numeric score $s_x \in [0, 1]$. Finally, the predicted values are thresholded at 0.5.

4.2 Domain-specific Forged Receipt Input

In order to better explore the semi-structured specific nature of the receipts, we experimented with four main types of input:

[1] The platform is available at `https://receipts.univ-lr.fr/`

1. **Text**: the raw text of a receipt without any pre-processing;
2. **Entities**: we detect the present entities based on a receipt ontology and concatenate them with a space separator (e.g. "Carrefour") as described below;
3. **Text + Entities**: we augment the receipt *Text* by introducing special markers for each type of entity (e.g., company, product) and replace each entity in the text with its text surrounded by the entity type markers [10];
4. **Knowledge-base Triples**: based on the same ontology but extracting also the semantic relations.

We, first, present the ontology, and then, we detail the detection and the usage of the entities and relations (triples).

Table 1. Receipt ontology object properties. Object properties connect two individuals (a subject and object) and can have a defined *domain* class to specify the class membership of the individuals serving as subjects, and an *image* class to define the class membership of the individuals serving as objects. The table does not list the *type* relation, associating every entity with its type. These data properties are the following: *has_date*, *has_time*, *amounts_to*, *has_total_price*, *has_number_of_items*, *has_full_payment_of*, *weights*, *has_price_per_kg*, *has_unit_price*, *has_quantity*, and *has_return_money_amount*.

Domain	Object Property	Inverse Property	Image
City	has_zipCode	is_zipCode_of	ZipCode
Company	has_contactDetail	is_contactDetail_of	ContactDetail
Company	has_address	is_address_of	Address
Company	has_email_address	is_email_address_of	EmailAdress
Company	has_fax	is_fax_of	FaxNumber
Company	has_website	is_website_of	Website
Company	has_phone_number	is_phone_number_of	PhoneNumber
Company	issued	is_issued_by	Receipt
Product	has_expansion	is_expansion_of	Expansion
Company	has_registration	is_registration_of	Registration
Receipt	has_intermediate_payment		IntermediatePayment
Receipt	concerns_purchase		Product
Receipt	contains	is_written_on	Product, Registration, ContactDetail
SIREN	includes	is_component_of	SIRET, RCS, TVA IntraCommunity
City, ZipCode	part_of		Address
Company	is_located_at		City
Company	sells	is_sold_by	Product

Receipt Ontology The receipt ontology was built by [5] and it is domain-specific, being curated to account for all the information present in the receipts (e.g., a concept for the receipts, instantiated by their IDs, a concept for the

issuing company, another for its contact information, etc.). The authors of the dataset chose to represent its semantic contents with an ontological model in order to explicitly represent what humans can imply through their understanding of an invoice. Indeed, much implicit information lies in the understanding of the layout, the format, the content and their combinations. For example, an address written under a company name corresponds to the address of the company. We can also note that a company can have several addresses, or that an address can be shared by several companies in the case of an office building. Moreover, the ontological model allows maintaining a certain flexibility (new classes can be added for other types of business documents, other types of registrations relevant to other countries, etc.) and enable the use of reasoners. However, we utilized the ontology as a starting point in our approach as we were not interested in the description, the inferences and the reasoning per se, but wanted to explore a less formal semantic enrichment of the content of the documents in order to propose a more generalizable approach. Therefore, we focused on the domain-specific entities and relations described and populated in the ontology to build our experiments. The entity detection and the knowledge-base triple detection are presented hereunder.

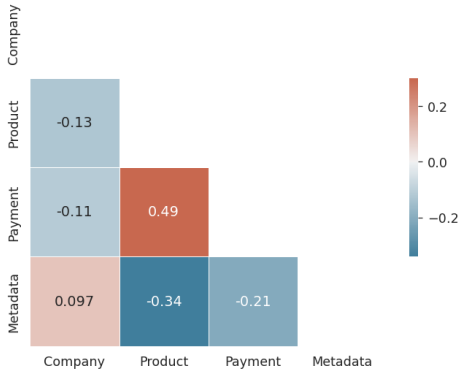
The ontology is originally written in French, consequently, all labels in this article have been translated, and it was automatically populated with manually-crafted regular expressions based on the regularities of a receipt document. For instance, the products and their prices were extracted from the lines of the document finished by the "€" symbol, or using it as a decimal separator, excluding the lines that report the total or the payment. The extraction process was performed as a finite state machine to adapt to more varied structures, such as prices and products not being aligned. The ontology was populated dynamically using the Python library Owlready2[2]. Table 1 lists the object properties that link the information present in the receipts in order to provide an exhaustive list of the extracted information. We note that the receipt is an entity itself, represented by the label of its ID (a numerical value).

Entity Detection We kept all domain-specific entity types, even in the cases where they produce redundant information, in order to maintain the granularity of the semantic annotation. For instance, when an address is correctly extracted, it is represented through several entity types: its full address, its city, and its postcode. Each entity subjected to any type of alteration (removal, addition, or modification) was counted. The modifications were not counted in themselves, only the entities altered were: for instance, a date "11/02/2017" altered to "10/02/2016" counts as one modified entity, even if it has suffered two graphic modifications. We grouped the entity types into four categories: company information (name, address, phone number, etc.), product information (label, price, quantity, or weight), payment information (total, paid amount, discount total), and receipt metadata (date, time, etc.). The number of modified entities per data split and entity types are presented in Table 4 (a).

[2] `https://owlready2.readthedocs.io/en/v0.37/`

Fig. 4. (a) Modified entities in the data splits. **(b)** Entity type correlation matrix.

Entity Type	Train	Test
Company	7	5
Product	33	44
Payment	30	37
Metadata	18	17
Total	88	103



Most of the altered entities involve amounts of money (product and payment entities), even if those are not always consistently modified. Figure 4 (b) shows how the forged receipts are correlated to the entity types. We notice a slightly high coefficient value (0.49) for payment and product entity types, proving a strong correlation between these two. This correlation proves the realistic nature of the forgeries, as an effort has been made to maintain the coherence of the receipt. Indeed, if the price of one or more products is modified, but the total amount remains as is, the forgery becomes easily detected by a human calculating the sum of the amounts. As these entities are not independent, we have considered these types of fraud as inconsistent, and consider them easier to detect by a context-based approach than forgeries involving the receipt metadata (i.e., date and time of purchase).

Finally, in order to take advantage of the semantic details of these entities (entity types), we modify the initial $\{x_i\}_{i=1}^{l}$ token sequence to give: $x = [x_0, x_1, \ldots, [ENTITY_{start}]x_i[ENTITY_{end}], \ldots, x_n]$ where n is the length of the sequence and $ENTITY$ is the entity type of $x_i \in [payment, company, etc.]$. We, afterward, feed this token sequence into $CamemBERT_{[CLS]}$ instead of $\{x_i\}_{i=1}^{l}$[3].

Knowledge-base Triple Detection In order to go beyond the extracted entities and provide more information about the relations between the entities, we chose to incorporate the domain-specific relationships present in the ontology curated by the authors of the dataset [5]. Our goal was to bring the underlying structure of the documents to the forefront by explicitly stating the relations between entities. Those relations include object properties, relations between entities such as *has_address* and *type* relations, that associate each entity with its class in the ontology. We made sure to remove inverse relations, e.g.,

[3] This strategy has been previously explored in research for different NLP tasks [35, 10, 11].

has_telephone_number and *is_telephone_number_of* by keeping only one of each pair. We also included attributive relations, i.e., data properties, that associate an entity with a value (numeric, date, or time). We use the extracted knowledge to render the semantic content of the receipts more explicit. Indeed, as the document's content does not have a syntactic structure, the extracted relations can help convey the underlying structure of the information present in the receipts. The knowledge-based triples serve as a text normalization of the content of the receipts, as a finite number of relations (object and data properties) describe all the information extracted.

5 Experiments

We compare our model with two baseline methods.

First, we consider a *numerical inconsistency checker*, by simulating a checker that assigns the forged class to any document in which there is a simple numerical inconsistency, not relying on any external knowledge. The numeric inconsistency checker accounts for forgeries that a human with a calculator could spot. We consider a simple numerical inconsistency any discrepancy between the total and the sum of the prices, between the total and the total paid, or between the quantity, the unitary price, and the price of the product. However, if the only numeric inconsistency lies in a tax estimation, we consider that our checker does not have the tools to notice the inconsistency, as it requires equation-solving skills.

Second, we consider a support vector machine (SVM) regression classifier with default hyperparameters as our baseline model applied on the term frequency-inverse document frequency (TF-IDF) representation of the unigrams and bigrams extracted from lowercased receipts.

We also compare our results to two CV methods proposed for the dataset. The Verdoliva [16, 15] architecture is also based on an SVM and combines three different approaches: a copy-move forgery detection module, based on [14], a camera-signature extraction (and comparison) module [16, 15], and a forgery detection based on local image features, originally proposed as a steganalysis method [13]. We also report the results proposed by Fabre [4], that fed the preprocessed images (discrete wavelet transform and grayscale) to a pre-trained model Resnet152 [23] for classification.

5.1 Hyperparameters

We experimented with an SVM with default hyperparameters (C of 1.0 and ϵ of 0.1). In the CamemBERT experiments, we use AdamW [26] with a learning rate of 1×10^{-5} for 2 epochs with mean squared error (MSE) loss. We also considered a maximum sentence length of 256 (no receipt is longer than this). We experiment with a CamemBERT endpoint (CamemBERT-base, with 110M parameters). The evaluation is performed in terms of precision (P), recall (R), and F1.

5.2 Results

Table 2 details the classification results. As the classification is very imbalanced, we report only the results for the "Forged" class.

Table 2. Evaluation results for forged receipt detection.

Method	**P**	**R**	**F1**
Numeric inconsistency checker	100.0	46.67	63.34
CV Approaches			
Fabre [4]	36.4	93.3	52.3
Verdoliva [4]	90.6	96.7	93.5
Baselines			
SVM (text)	7.73	53.33	13.50
SVM (entities)	5.24	33.33	9.05
SVM (text + entities)	5.77	40.00	10.08
SVM (triples)	**29.41**	**100.0**	**45.45**
CamemBERT Approaches			
CamemBERT (text)	6.61	50.0	11.67
CamemBERT (entities)	8.76	73.33	15.66
CamemBERT (text + entities)	7.39	63.33	13.24
CamemBERT (triples)	93.75	**100.0**	**96.77**

We notice how the methods using *Triples* as their input outperform the others, even in their TF-IDF representation, recall is equal to one, meaning all forged receipts are successfully retrieved. Figure 5 presents the area under the receiver operating characteristic (ROC) curve (AUC) for our CamemBERT-based experiments. Not surprisingly, we observe that the *Triples* approach has an AUC near to 1 which means it has a good measure of separability, while the others are closer to 0.

In the case of *Triples*, we observed only two mislabelled true receipts. In one of them, receipt 211, the total price is rather blurry in the image, so it has been manually corrected in the OCR output. However, the value of the total uses "," instead of "." as a decimal separator. As we can see in Figure 1, the usual separator is ".". This manually induced irregularity could explain this error. The other mislabelled true receipt shows no salient irregularity. The only thing we have noted is that the total amount is expensive (over 87 euros).

Concerning the comparison with the numeric inconsistency checker, we noted that our approach performs better, even for receipts without numerical inconsistency. However, it is important to take notice of the strict definition of inconsistency we have used. Indeed, we only consider inconsistencies in the interaction of the entities themselves, as it provides a stable way to annotate. However, most of the "consistent" forgeries that we consider the most difficult to detect (and the numeric inconsistency checker misses) are implausible. For instance, many of the receipts in which only the date has been modified are actually assigned to an impossible date as in receipt number 334, where the month is numbered 17

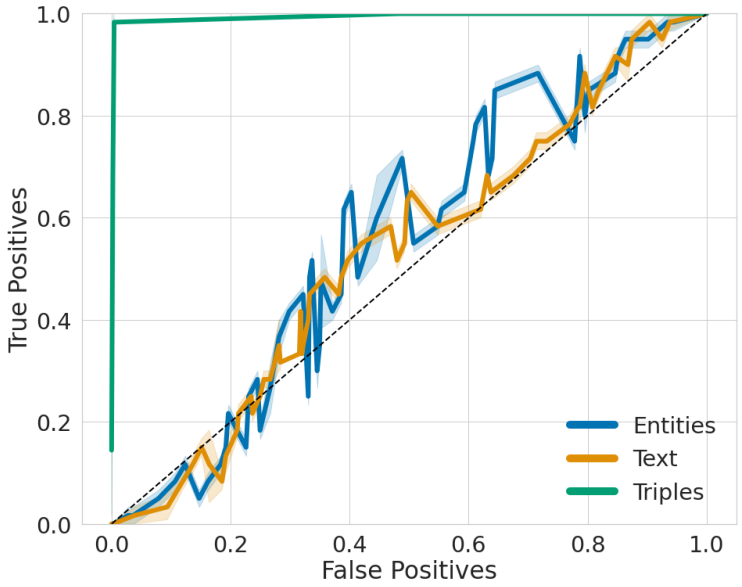


Fig. 5. ROC curve.

or receipt number 662, where the year (2018) is actually after the data collection stopped.

Results per Entity Types We analysed the results in terms of their count of modified entities and found that it has no statistical impact. We performed an independent t-test in order to compare the number of altered entities in correctly detected forged receipts with the number of altered entities in undetected forged receipts with the results of our approaches based on *Text*, *Entities* and *Text + Entities*. We did not find any statistically significant difference in the means of the two groups, whether we looked into the count of total modified entities, the count of product-related entities, company-related, payment-related, or the receipt metadata entities (p-value $> 5\%$). There was no use in analysing this kind of error for the triples approach (a recall of 100%).

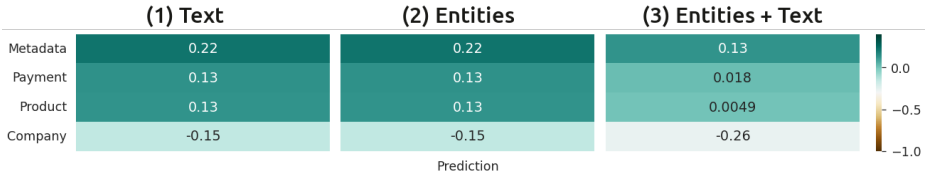


Fig. 6. Entity types correlating with the fraud predictions.

Figure 6 presents the correlation coefficients between the predictions of each model and the modified entity types. We observe for all approaches a rather weak positive correlation between the existence of forged Metadata, Payment, and Product entities, and an even weaker negative correlation with the Company-related entities. This allows us to understand that, while there is an influence regarding which entities are forged, generally, the Metadata, Payment, and Product entities could be correlated with the performance of our fraud detection methods.

Results per Ontology Relation Types We also analysed the results of forged receipt detection using only one relation type at a time, as shown in Table 3. The relation *contains* exists between the receipt ID and any registration or contact detail of the company present in the receipt (phone, email, fax, address, etc.). Keeping the *contains* relation allows to keep and structure the information related to the company-specific entities. The model trained with such input mislabels uncommon receipts, such as a highway toll or an hourly parking ticket, whose structure is very different from supermarket receipts. Company information is not among the most modified (only seven entities in the train set and five in the test), which leads to the belief that the data may have other biases. For instance, up to almost 50% of the forged receipts were emitted by the same company (Carrefour). Carrefour is indeed the most common company, but it represents only 30% of all receipts. Moreover, the ID associated with each receipt is not entirely random, as receipts are at least sorted by their emitting company.

Table 3. Evaluation results for forged receipt detection.

Method	P	R	F1
Per triple type			
amounts_to	63.83	100.0	77.92
contains	44.12	100.0	61.22

Furthermore, *amounts_to* is the relation to the value of the amount of the intermediate payment. When there is only one mean of payment, it is equivalent to the total amount, however, when two or more means of payment are used, *amounts_to* projects the relation to those amounts. As we can see in Table 3, keeping only this relation still yields very effective results. This triple type considers exclusively information related to general numeric values (totals and payment information). A certain bias is to be expected in the modified numerical values, such as Benford's law [37] which describes the non-normal distribution of naturally occurring numerical data, and it has been used in accounting fraud detection. In real-life occurring numbers (such as prices, population numbers, etc.), the first digit is likely to be small. Indeed, the authors of the dataset report that using Benford's law to look for anomalous numerical data results in a recall of

70% [5]. These results, taken as input only the triples *amounts_to*, are very encouraging for the ability of our approach to leverage statistical information, even on numerical values, to detect forgery.

6 Conclusions and Future Work

This paper proves that content-based methods are up to the challenge of document fraud detection on the same level as image-based methods. Our initial goal was to build a baseline and to encourage future work in the NLP domain to address document forgery detection, and the results exceeded our expectations. Our semantic-aware approach based on the CamemBERT pre-trained model projecting the relations between entities to represent the content of the receipts achieves high recall values by efficiently leveraging the information extracted from the documents in the form of triples. Ideally, we would like to test our approach on other realistic forgery datasets in order to experiment with other document types and more complex use-cases, however we know of no other publicly available forgery detection corpus. Moreover, as administrative documents are often exchanged as their scanned images, as future work, we propose to continue this line of work by using multimodal approaches [29, 48, 47].

Acknowledgements

This work was supported by the French defence innovation agency (AID), the VERINDOC project funded by the Nouvelle-Aquitaine Region.

References

1. Abramova, S., et al.: Detecting copy–move forgeries in scanned text documents. Electronic Imaging **2016**(8), 1–9 (2016)
2. Ahmed, A.G.H., Shafait, F.: Forgery detection based on intrinsic document contents. In: 2014 11th IAPR International Workshop on Document Analysis Systems. pp. 252–256 (2014)
3. Artaud, C., Doucet, A., Ogier, J.M., d'Andecy, V.P.: Receipt dataset for fraud detection. In: First International Workshop on Computational Document Forensics (2017)
4. Artaud, C., Sidère, N., Doucet, A., Ogier, J.M., Yooz, V.P.D.: Find it! Fraud detection contest report. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 13–18 (2018)
5. Artaud, C.: 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 (2019)
6. Behera, T.K., Panigrahi, S.: Credit card fraud detection: a hybrid approach using fuzzy clustering & neural network. In: 2015 Second International Conference on Advances in Computing and Communication Engineering (2015)

7. Benchaji, I., Douzi, S., El Ouahidi, B.: Using genetic algorithm to improve classification of imbalanced datasets for credit card fraud detection. In: International Conference on Advanced Information Technology, Services and Systems (2018)
8. Bertrand, R., Gomez-Krämer, P., Terrades, O.R., Franco, P., Ogier, J.M.: A system based on intrinsic features for fraudulent document detection. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 106–110. Washington, DC, USA (Aug 2013)
9. Bertrand, R., Terrades, O.R., Gomez-Krämer, P., Franco, P., Ogier, J.M.: A conditional random field model for font forgery detection. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 576–580 (2015)
10. Boros, E., Moreno, J., Doucet, A.: Event detection with entity markers. In: European Conference on Information Retrieval. pp. 233–240 (2021)
11. Boros, E., Moreno, J.G., Doucet, A.: Exploring entities in event detection as question answering. In: European Conference on Information Retrieval. pp. 65–79. Springer (2022)
12. Carta, S., Fenu, G., Recupero, D.R., Saia, R.: Fraud detection for e-commerce transactions by employing a prudential multiple consensus model. Journal of Information Security and Applications **46** (2019)
13. Cozzolino, D., Gragnaniello, D., Verdoliva, L.: Image forgery detection through residual-based local descriptors and block-matching. In: 2014 IEEE International Conference on Image Processing (ICIP) (2014)
14. Cozzolino, D., Poggi, G., Verdoliva, L.: Efficient dense-field copy–move forgery detection. IEEE Transactions on Information Forensics and Security **10**(11) (2015)
15. Cozzolino, D., Verdoliva, L.: Camera-based image forgery localization using convolutional neural networks. In: 2018 26th European Signal Processing Conference (EUSIPCO) (2018)
16. Cozzolino, D., Verdoliva, L.: Noiseprint: A cnn-based camera model fingerprint. IEEE Transactions on Information Forensics and Security **15**, 144–159 (2020)
17. Cruz, F., Sidere, N., Coustaty, M., d'Andecy, V.P., Ogier, J.M.: Local binary patterns for document forgery detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1 (2017)
18. Cruz, F., Sidère, N., Coustaty, M., Poulain D'Andecy, V., Ogier, J.: 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. pp. 117–124 (2018)
19. Elkasrawi, S., Shafait, F.: Printer identification using supervised learning for document forgery detection. In: 2014 11th IAPR International Workshop on Document Analysis Systems. pp. 146–150 (2014)
20. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security **7**(3) (2012)
21. Gomez-Krämer, P.: Verifying document integrity. Multimedia Security 2: Biometrics, Video Surveillance and Multimedia Encryption pp. 59–89 (2022)
22. Guo, H., Yuan, S., Wu, X.: Logbert: Log anomaly detection via bert. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2021)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015). https://doi.org/10.48550/ARXIV.1512.03385, `https://arxiv.org/abs/1512.03385`
24. James, H., Gupta, O., Raviv, D.: Ocr graph features for manipulation detection in documents (2020)
25. Kim, J., Kim, H.J., Kim, H.: Fraud detection for job placement using hierarchical clusters-based deep neural networks. Applied Intelligence **49**(8) (2019)

26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
27. Kowshalya, G., Nandhini, M.: Predicting fraudulent claims in automobile insurance. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (2018)
28. Lee, Y., Kim, J., Kang, P.: Lanobert: System log anomaly detection based on bert masked language model. arXiv preprint arXiv:2111.09564 (2021)
29. Li, P., Gu, J., Kuen, J., Morariu, V.I., Zhao, H., Jain, R., Manjunatha, V., Liu, H.: Selfdoc: Self-supervised document representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5652–5660 (2021)
30. Li, Y., Yan, C., Liu, W., Li, M.: 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) (2016)
31. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. ArXiv **abs/1907.11692** (2019)
32. Martin, L., Muller, B., Ortiz Suárez, P.J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7203–7219. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.645, `https://aclanthology.org/2020.acl-main.645`
33. Mikkilineni, A.K., Chiang, P.J., Ali, G.N., Chiu, G.T., Allebach, J.P., Delp III, E.J.: Printer identification based on graylevel co-occurrence features for security and forensic applications. In: Security, Steganography, and Watermarking of Multimedia Contents VII. vol. 5681, pp. 430–440. International Society for Optics and Photonics (2005)
34. Mishra, A., Ghorpade, C.: 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) (2018)
35. Moreno, J.G., Boros, E., Doucet, A.: Tlr at the ntcir-15 finnum-2 task: Improving text classifiers for numeral attachment in financial social data. In: Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo Japan. pp. 8–11 (2020)
36. Nadim, A.H., Sayem, I.M., Mutsuddy, A., Chowdhury, M.S.: Analysis of machine learning techniques for credit card fraud detection. In: 2019 International Conference on Machine Learning and Data Engineering (iCMLDE). pp. 42–47 (2019)
37. Nigrini, M.J.: Benford's Law: Applications for forensic accounting, auditing, and fraud detection, vol. 586. John Wiley & Sons (2012)
38. Rabah, C.B., Coatrieux, G., Abdelfattah, R.: The supatlantique scanned documents database for digital image forensics purposes. In: 2020 IEEE International Conference on Image Processing (ICIP) (2020)
39. Rizki, A.A., Surjandari, I., Wayasti, R.A.: Data mining application to detect financial fraud in indonesia's public companies. In: 2017 3rd International Conference on Science in Information Technology (ICSITech) (2017)
40. Rossi, A., Firmani, D., Matinata, A., Merialdo, P., Barbosa, D.: Knowledge Graph Embedding for Link Prediction: A Comparative Analysis. ACM Trans. Knowl. Discov. Data **15**(2), 14:1–14:49 (2021)

41. Shang, S., Kong, X., You, X.: Document forgery detection using distortion mutation of geometric parameters in characters. Journal of Electronic Imaging **24**(2), 023008 (2015)
42. Sidere, N., Cruz, F., Coustaty, M., Ogier, J.M.: A dataset for forgery detection and spotting in document images. In: 2017 Seventh International Conference on Emerging Security Technologies (EST) (2017)
43. Tornés, B.M., Boros, E., Doucet, A., Gomez-Krämer, P., Ogier, J.M., d'Andecy, V.P.: Knowledge-based techniques for document fraud detection: A comprehensive study. In: Computational Linguistics and Intelligent Text Processing: 20th International Conference, CICLing 2019, La Rochelle, France, April 7–13, 2019, Revised Selected Papers, Part I. pp. 17–33. Springer (2023)
44. Van Beusekom, J., Shafait, F., Breuel, T.M.: Text-line examination for document forgery detection. International Journal on Document Analysis and Recognition (IJDAR) **16**(2), 189–207 (2013)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
46. Vidros, S., Kolias, C., Kambourakis, G., Akoglu, L.: Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. Future Internet **9**(1) (2017)
47. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In: ACL-IJCNLP 2021 (2021)
48. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (2020)