RESEARCH ARTICLE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 09 September
www.rspsciencehub.com

http://dx.doi.org/10.47392/IRJASH.2023.064

AI Framework for Non-Native PDF Comparison

Maragathamani Subramaniam [1], Diwakar Ojha [2], Anshuman Mahapatra [3], Subhashini Lakshminarayan [4]
[1]Data Science Senior Analyst Accenture, Coimbatore, Tamilnadu, India
[2]Data Scientist, Accenture, Coimbatore, Tamilnadu, India
[3]Data Science Senior Manager, Accenture, Coimbatore, Tamilnadu, India
[4]Data Science Senior Manager , Accenture, Coimbatore, Tamilnadu, India

Emails: m.m.subramani@accenture.com, dojha00@gmail.com, anshuman.a.mahapatra@accenture.com, s.j.lakshminarayanan@accenture.com

Article History

Received: 14 August 2023
Accepted: 28 September 2023
Published: 30 September 2023

Keywords:

Layout Detection;
PDF com parison;
PDF Page labelling;
Image change tracking

Abstract

Figuring out how a document has changed from one version to another isn't always the simplest task. We encountered the problem of comparing two PDF documents, edited using different editing tools. When we tried to compare these PDFs, using existing comparison tools, comparison results were not satisfactory. After analysis, we found that, if documents had been edited using any other tool than acrobat(non-Native), then these tools were unable to detect the proper layout (para, header, footer, columns, tables etc.) of the document and therefore unable to sequence them in correct order resulting in false comparison output. To overcome this problem, we tried latest developments in computer vision to detect the layout information of the document. Using layout information, contents were arranged in correct order and then compared. This resulted in better comparison output. Also, using AI for layout detection made it independent of how the document was created and edited. We built a complete framework which includes reading the information, detecting layout, arranging information, comparing it, and visualizing the differences. This Framework can be applied to build any document comparison tool irrespective of document type.

1. Introduction

Documents are created to preserve content in the easiest manner. One of the greatest advances in the digitized era is to store vast amounts of data electronically. (Smock, Pesala, and Abraham) The dimensions of electronic data are huge than paper documents. Electronic documents in place of physical documents save cost reduction, storage space, portability, zero damage and standard structure which in turn results in easy access. It also provides efficient ways to store and retrieve information. PDF is one of the predominantly used document formats to store textual contents in the organizations. (Kardas et al.) It provides multiple functionalities to users like search, index, images etc. In addition to it, there is also facility to edit and add text or images, inside existing PDF. A PDF can contain information in formats like Text, Table, and Image data. Due to wide acceptance of PDFs and functionality of editing, many times we face challenge of finding out the differences between original and edited version of the PDF. There are many tools available in market which perform these differences and highlight them. (Gemelli, Vivoli, and

Marinai) But we found these tools works well when source pf PDF document is acrobat but when PDF are created or edited from other tools then these comparison utilities are unable to align the content in proper order and hence their comparison does not show correct differences. And this comparison becomes more incorrect if PDFs are written in two columnar formats. To overcome this challenge, we proposed to use advanced ML model to detect the PDF. (Bimbo et al.)

Layout and then arrange the information using this layout information. This Machine learning model is trained on publicly available PDF documents. It takes PDF page as input in format of image and provides the bounding boxes for each columnar text data in the PDF. Our emphasis is to increase the PDF comparison accuracy with advances in AI and current technology.

2. Related work

There are multiple tools available online as paid services or as opensource for comparison utility.

2.1. Paid Comparison tools

Some of the paid comparison tools are Beyond Compare, Foxit Compare (Xiong and Foxit), Adobe Compare etc. They work great in case of standard PDFs. But we found that if PDFs are edited in non-standard way or PDFs are generated from different sources other than Acrobat reader, they are not able to align the information from the PDF which leads to highlight incorrect differences.

2.2. Open-Source Tools

There are very few compare utilities available where comparison result is good enough to use them. One of them is pdf-diff (Wu et al.) utility created by Josh-Data on github, this utility is created using python library and works great if PDFs are of standard form and in one columnar format but does not work on non- standard and modified PDF's. Also, it does not perform image comparison. It also shows differences in text only and does not show differences of space or new lines characters added and removed in PDF.

Some of the other available tools are Diff compare, Draftable, Pdfforge, Kiwi pdf compare. (Jarvis)

Diff compare highlights differences in side-by-side view but fails in terms of accuracy compared with other ones.

Draftable compares pdf side-by-side view and spot the differences based on style and content. It works on text and images but fails in case PDFs are edited using any other tool than acrobat.

Pdf forge is a platform to edit, create, convert, and organize PDFs. It compares two PDFs based on text changes. It offers side-side or inline view. It doesn't work for images.

Kiwi pdf compare has text and image compare. On account of images, pixel to pixel comparison comes into play. It works on the pdf from other sources too. However, the free version has limits, you can compare up to 100 pages. (Islam, Dias, and Sunda-Meya)

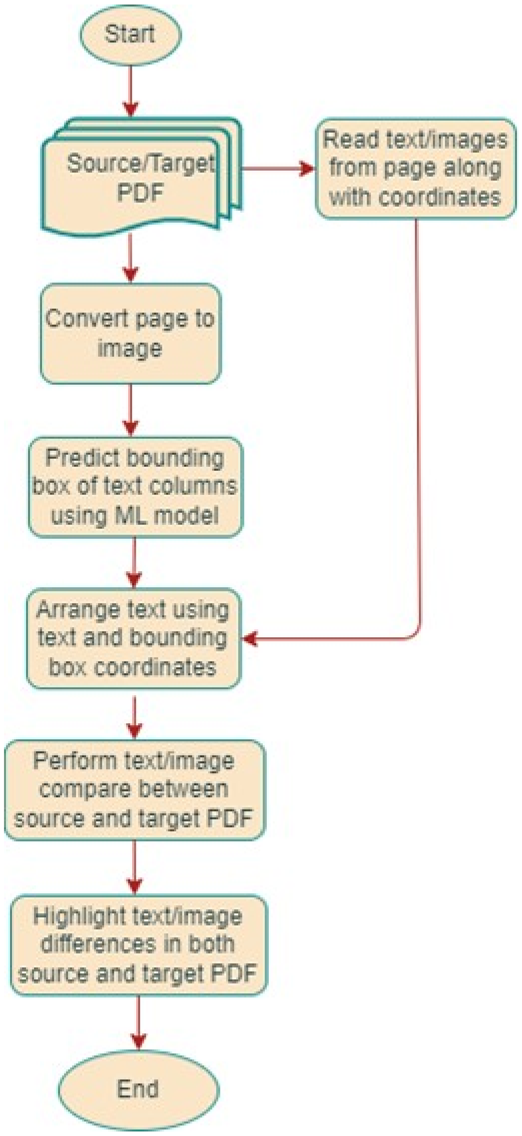
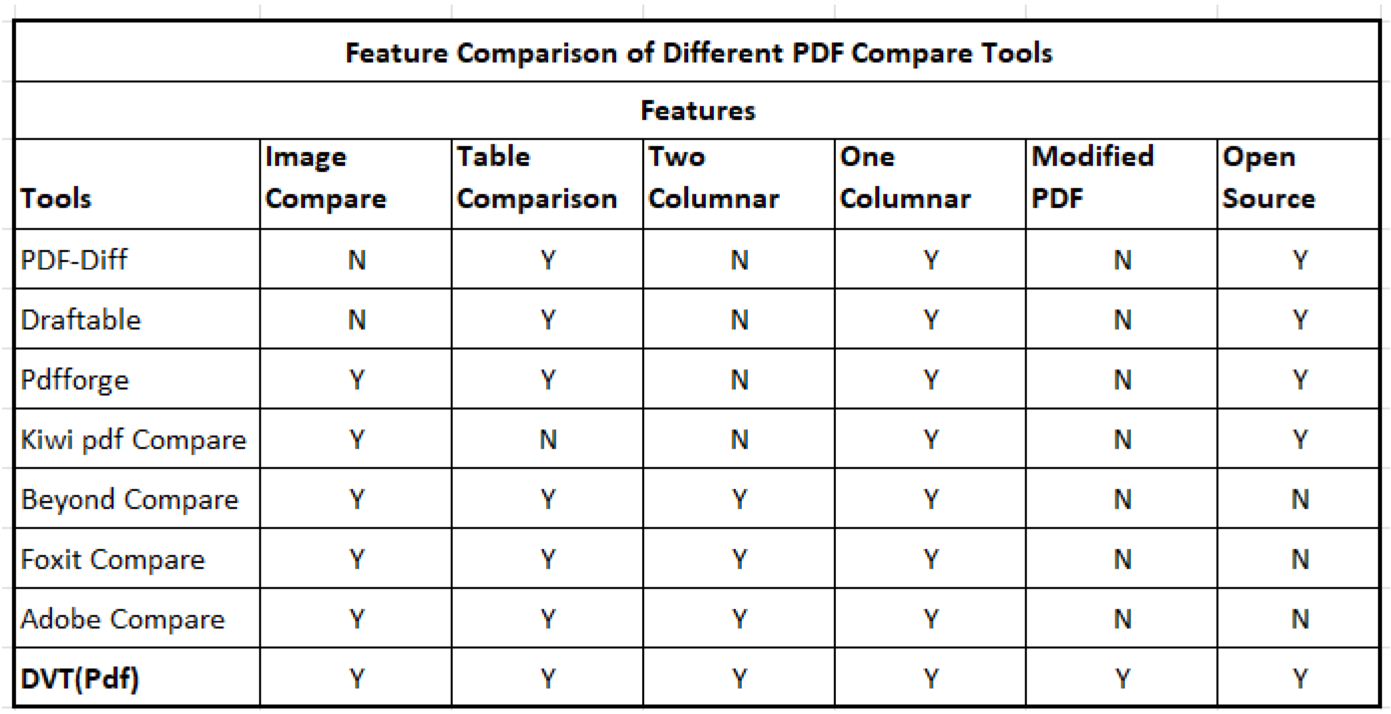

Start
Source/Target PDF
Read text/images from page along with coordinates
Convert page to image
Predict bounding box of text columns using ML model
Arrange text using text and bounding box coordinates
Perform text/image compare between source and target PDF
Highlight text/image differences in both source and target PDF
End

FIGURE 1. **Process flow diagram**



Feature Comparison of Different PDF Compare Tools						
	Features					
Tools	**Image Compare**	**Table Comparison**	**Two Columnar**	**One Columnar**	**Modified PDF**	**Open Source**
PDF-Diff	N	Y	N	Y	N	Y
Draftable	N	Y	N	Y	N	Y
Pdfforge	Y	Y	N	Y	N	Y
Kiwi pdf Compare	Y	N	N	Y	N	Y
Beyond Compare	Y	Y	Y	Y	N	N
Foxit Compare	Y	Y	Y	Y	N	N
Adobe Compare	Y	Y	Y	Y	N	N
DVT(Pdf)	Y	Y	Y	Y	Y	Y

FIGURE 2. Feature comparison of different PDF compare tools.

3. Proposed approach

In this section we will present complete framework created for comparing documents. This framework has been created for comparing PDF documents but can be applied to other document types also. It uses state of the art computer vision ML capability to align texts in PDFs and then perform compare. It also captures images from PDFs and compares their position and content to verify if they have not been changed. Once Source and Target PDFs have been supplied, it processes them page by page and fetches text/images from both (source and target) the PDFs along with their respective coordinate in the page. Also converts the pages to image format and sends them to ML model to get bounding boxes for the text columns. Once text coordinates and bounding box coordinates are available, text are aligned as per their position. Also, if we find texts are overlapping row wise within same column then considers them in same line. After texts are arranged for same page in both PDFs, we use text compare utility to compare the texts and find out differences in text, which includes difference in text and tables. Now we compare images available on both the pages for their position and content and if any changes are found then it is marked as different. Once differences are found, we use our existing text coordinate data and find out coordinates of the differences and then use PyMuPDF (Tkachenko et al.) pdf highlight functionality to highlight these differences in PDF format. If output is requested in image format, pages are converted to image and differences are highlighted using PIL. Below is the process flow diagram of the tasks performed to create this utility.

Below are the five steps framework to perform the PDF compare:

1. Extract text and Images from Both PDFs
2. Train Model to detect Layout information.
3. Align Texts as per layout information.
4. Compare text and image differences.
5. Highlight differences.

A. Extract text and Images from PDFs

We are using existing Python libraries to fetch Text and Images from the PDF files. Text information is captured at character level along with the position on the page. Position is captured as a distance from left-top corner as coordinates. New line and space character information is also captured.

3.1. Train Model to detect Layout information

Once character information is captured, we need to realign the text information if they are not sequentially aligned. To align the information, we need to know if the PDF page is in one columnar format or two columnar format. If information is present in two columnar format, then what is the column coordinates so that while aligning the text, we can divide the text in different columns and then align it, to get properly sequenced data. To capture PDF text alignment, we have trained Detectron2 (Wu et al.) model which takes PDF page converted to jpeg image as input and returns Bounding box information of text columns in two columnar PDF format. Below are the steps performed to train the Detectron2 model for detecting PDF layouts:

1) Data Collection: As part of data collection, we down- loaded publicly available two columnar PDF documents. It covered variety of two columnar PDF pages. These pages were converted to image format.

2) Data Labeling: We used label studio (Tkachenko et al.) to create bounding box around text columns for two columnar pages. This labeled information is exported in Detectron2 input format.

3) Model Training: We used Detectron2 model from Facebook and performed transfer learning to train on creating bounding boxes around text columns in case of two columnar PDF.

4) Bounding Box Prediction: Page images are sent to our fine-tuned model which in turn returns bounding box information for each column.

3.2. Align Texts as per layout information

Once layout information is available, texts are aligned as per the bounding box coordinates. Also, if texts are found to be overlapping with 60% are greater overlap ratio, then we align them in same rows.

3.3. Compare text and image differences

Once texts are aligned, we are using Python library from Google(diff-match-patch) (Cross et al. Marinai) to compare this text information from the pages. Also, images and their positions are compared, and if any difference arises either in image position or image content then this image is highlighted.

3.4. Highlight differences

Once text and image differences are available, we highlight these differences in the easily comprehensible format. Output of PDF compare can be seen in image 4.

4. Experimental results

4.1. Use rule-based system to compare PDF.

First simple rule-based model was used to arrange text in a PDF document. Texts are first arranged from top to bottom and then from left to right, this method is used by many PDF readers to align texts. it performs well on PDFs having only one columnar text. But in case of two columnar format PDFs, finding column separation and using that information to arrange text data was not possible since pages can contain images, tables etc. in same page. So, we used AI bounding box prediction to find columns bounding boxes.

We also tried using OpenCV to identify coordinates of columns, but different PDFs have different way to organize.

the content. In some PDFs, tables are spanned across both the columns and take full width of the page. It is difficult to capture columns coordinates using OpenCV in such scenarios. We found many other formats where capturing layout details using OpenCV was difficult. So, we decided to train ML model for it.

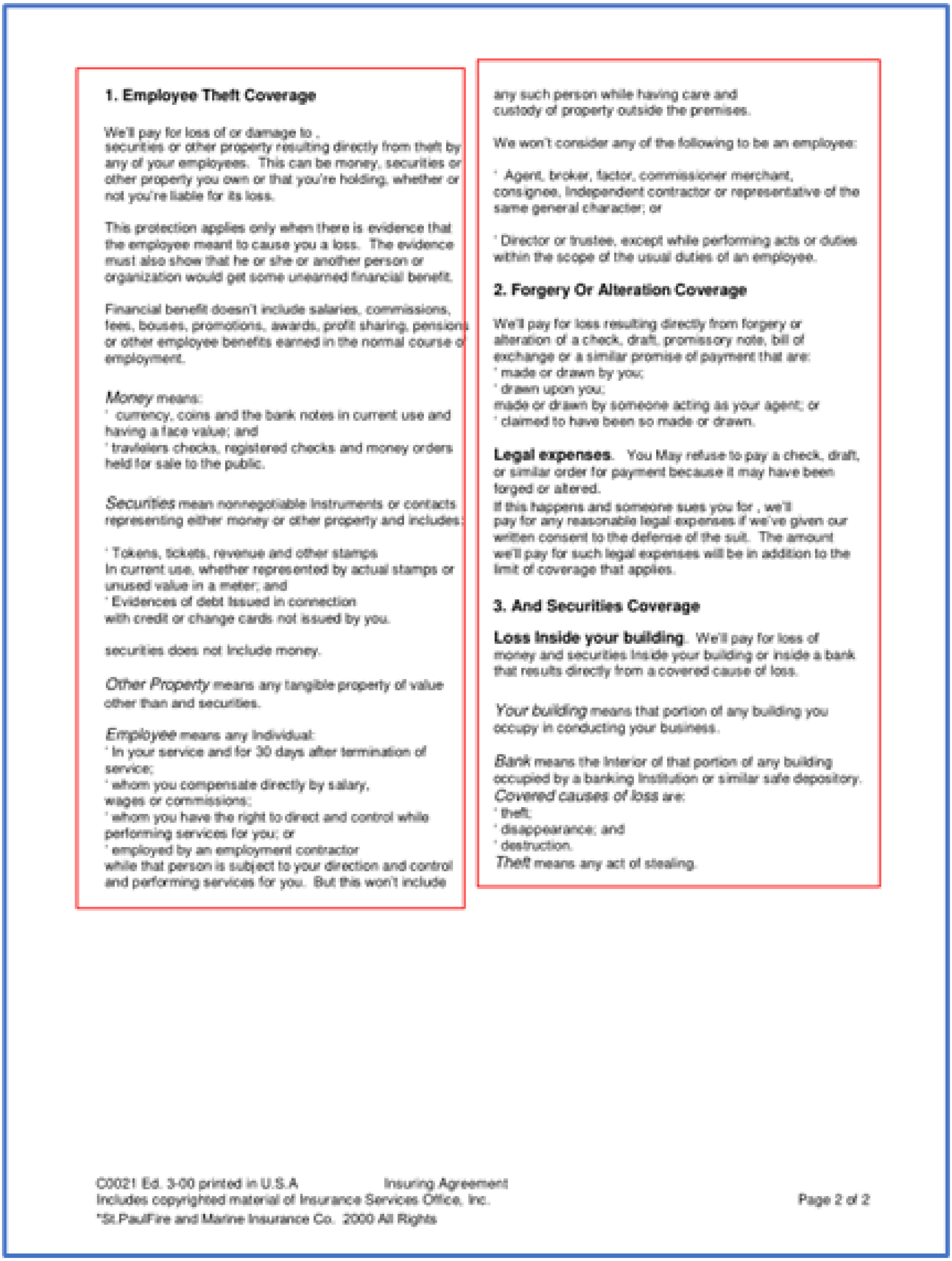
4.2. Build model to predict bounding box and compare.

- **Model training with masked data:** We tried to mask the text data using OpenCV (&apos;) and train Detectron2 model to detect text column bounding boxes, but it doesn't perform well in terms of accuracy.

- **Model training without masking data:** We tried to label unmasked PDF pages using LabelStudio and then trained Detectron2 model over it. Models gave better results in detecting bounding boxes for textual columns.

5. Conclusion and future work

In this paper we have presented a framework using AI to create documents and compare utility for PDFs. We found differences in text, images, tables, numbers, or symbols in PDFs. Comparison is done in quick time with greater accuracy. It is also independent of how document was edited. We can also use this framework to create compare utility for other document types. This will help people in multiple domains where PDF comparison is required. This tool can be useful for document review in organizations, and it can present the differences in easy to understandable format between different versions



1. Employee Theft Coverage

We'll pay for loss of or damage to , securities or other property resulting directly from theft by any of your employees. This can be money, securities or other property you own or that you're holding, whether or not you're liable for its loss.

This protection applies only when there is evidence that the employee meant to cause you a loss. The evidence must also show that he or she or another person or organization would get some unearned financial benefit.

Financial benefit doesn't include salaries, commissions, fees, bouses, promotions, awards, profit sharing, pensions or other employee benefits earned in the normal course of employment.

Money means:

' currency, coins and the bank notes in current use and having a face value; and

' travlelers checks, registered checks and money orders held for sale to the public.

Securities mean nonnegotiable Instruments or contacts representing either money or other property and includes:

' Tokens, tickets, revenue and other stamps In current use, whether represented by actual stamps or unused value in a meter; and

' Evidences of debt Issued in connection with credit or change cards not issued by you.

securities does not Include money.

Other Property means any tangible property of value other than and securities.

Employee means any Individual:

' In your service and for 30 days after termination of service;

' whom you compensate directly by salary, wages or commissions;

' whom you have the right to direct and control while performing services for you; or

' employed by an employment contractor while that person is subject to your direction and control and performing services for you. But this won't include any such person while having care and custody of property outside the premises.

We won't consider any of the following to be an employee:

' Agent, broker, factor, commissioner merchant, consignee, Independent contractor or representative of the same general character; or

' Director or trustee, except while performing acts or duties within the scope of the usual duties of an employee.

2. Forgery Or Alteration Coverage

We'll pay for loss resulting directly from forgery or alteration of a check, draft, promissory note, bill of exchange or a similar promise of payment that are:

' made or drawn by you;

' drawn upon you;

made or drawn by someone acting as your agent; or

' claimed to have been so made or drawn.

Legal expenses. You May refuse to pay a check, draft, or similar order for payment because it may have been forged or altered.

If this happens and someone sues you for , we'll pay for any reasonable legal expenses if we've given our written consent to the defense of the suit. The amount we'll pay for such legal expenses will be in addition to the limit of coverage that applies.

3. And Securities Coverage

Loss Inside your building. We'll pay for loss of money and securities Inside your building or inside a bank that results directly from a covered cause of loss.

Your building means that portion of any building you occupy in conducting your business.

Bank means the Interior of that portion of any building occupied by a banking Institution or similar safe depository.

Covered causes of loss are:

' theft;

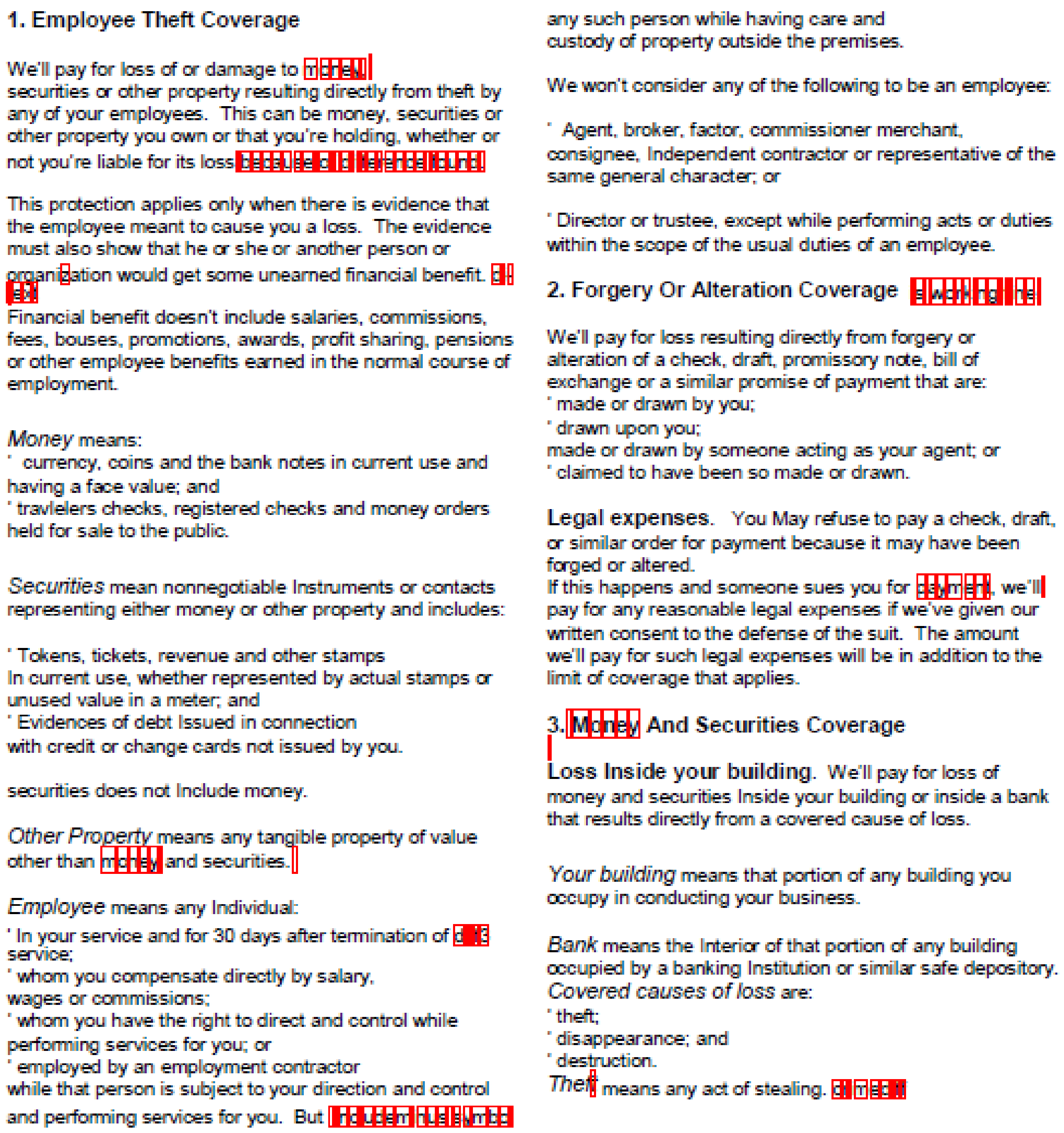
' disappearance; and

' destruction.

Theft means any act of stealing.

C0021 Ed. 3-00 printed in U.S.A Insuring Agreement
Includes copyrighted material of Insurance Services Office, Inc. Page 2 of 2
"St.PaulFire and Marine Insurance Co. 2000 All Rights

FIGURE 3. Column box labelling of PDF page



1. Employee Theft Coverage

We'll pay for loss of or damage to money, securities or other property resulting directly from theft by any of your employees. This can be money, securities or other property you own or that you're holding, whether or not you're liable for its loss

This protection applies only when there is evidence that the employee meant to cause you a loss. The evidence must also show that he or she or another person or organization would get some unearned financial benefit.

Financial benefit doesn't include salaries, commissions, fees, bouses, promotions, awards, profit sharing, pensions or other employee benefits earned in the normal course of employment.

Money means:
' currency, coins and the bank notes in current use and having a face value; and
' travlelers checks, registered checks and money orders held for sale to the public.

Securities mean nonnegotiable Instruments or contacts representing either money or other property and includes:

' Tokens, tickets, revenue and other stamps
In current use, whether represented by actual stamps or unused value in a meter; and
' Evidences of debt Issued in connection
with credit or change cards not issued by you.

securities does not Include money.

Other Property means any tangible property of value other than money and securities.

Employee means any Individual:
' In your service and for 30 days after termination of
service;
' whom you compensate directly by salary, wages or commissions;
' whom you have the right to direct and control while performing services for you; or
' employed by an employment contractor
while that person is subject to your direction and control and performing services for you. But

any such person while having care and custody of property outside the premises.

We won't consider any of the following to be an employee:

' Agent, broker, factor, commissioner merchant, consignee, Independent contractor or representative of the same general character; or

' Director or trustee, except while performing acts or duties within the scope of the usual duties of an employee.

2. Forgery Or Alteration Coverage

We'll pay for loss resulting directly from forgery or alteration of a check, draft, promissory note, bill of exchange or a similar promise of payment that are:
' made or drawn by you;
' drawn upon you;
made or drawn by someone acting as your agent; or
' claimed to have been so made or drawn.

Legal expenses. You May refuse to pay a check, draft, or similar order for payment because it may have been forged or altered.
If this happens and someone sues you for payment, we'll pay for any reasonable legal expenses if we've given our written consent to the defense of the suit. The amount we'll pay for such legal expenses will be in addition to the limit of coverage that applies.

3. Money And Securities Coverage

Loss Inside your building. We'll pay for loss of money and securities Inside your building or inside a bank that results directly from a covered cause of loss.

Your building means that portion of any building you occupy in conducting your business.

Bank means the Interior of that portion of any building occupied by a banking Institution or similar safe depository.
Covered causes of loss are:
' theft;
' disappearance; and
' destruction.
Theft means any act of stealing.

FIGURE 4. Bounding box prediction for page with images

of same document. For academics this serves as a useful tool to get extra information handy without thoroughly reading through the document. In review process, this will also enhance the accuracy of comparison. This tool also overcomes the limitations set by the open-source tools in the market like image comparison, table comparison and deals with two columnar PDFs created from different sources.

We can train and replace the model to deal with more complex PDF files. Also, this framework can be extended to other document types like docx, pptx, etc.,

References

&apos; Wscg. “The 26th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2016 in co-operation with EUROGRAPHICS”. 2018. 13–18.

Bimbo, Del, et al. “Data Augmentation on Graphs for Table Type Classification”. *Structural, Syntactic, and Statistical Pattern Recognition. S+SSPR 2022* 13813 (2022).

Gemelli, Andrea, Emanuele Vivoli, and Simone Marinai. “Graph Neural Networks and Representation Embedding for Table Extraction in PDF Documents”. *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022. 1719–1726.

Islam, Shafiqul, Jorge Dias, and Anderson Sunda-Meya. “On the Design and Development of Vision-Based Autonomous Mobile Manipulation”. *IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2021. 1–6.

Jarvis, R A. “A Perspective on Range Finding Techniques for Computer Vision”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-5.2 (1983): 122–139.

Kardas, Marcin, et al. “AxCell: Automatic Extraction of Results from Machine Learning Papers”. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020).

Smock, Brandon, Rohith Pesala, and Robin Abraham. “PubTables-1M: Towards comprehensive table extraction from unstructured documents”. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022): 4624–4632.

Embargo period: The article has no embargo period.

To cite this Article: , Maragathamani Subramaniam, Diwakar Ojha , Anshuman Mahapatra , and Subhashini Lakshminarayan . “AI Framework for Non-Native PDF Comparison.” International Research Journal on Advanced Science Hub 05.09 September (2023): 358–364. http://dx.doi.org/10.47392/IRJASH.2023.064