Automation of Information Extraction from Financial Statements using Graph-Based Techniques

Data Science Division – Anurag Bejju, Saeid Malladavoudi, Monica Pickard

Objective

- To correctly **identify** and **extract** key variables (eg: *total assets*) from the correct table (eg: *Balance Sheet*) in an annual financial statement pdf document of a company.

Scope

- Extract '*total revenue*' from income statement.
- Extract '*total assets*' from consolidated statement of financial position.
- Extract the '*currency*' used in the consolidated financial statements.
- Identify the pages that contain '*geographical segmentation*' information.

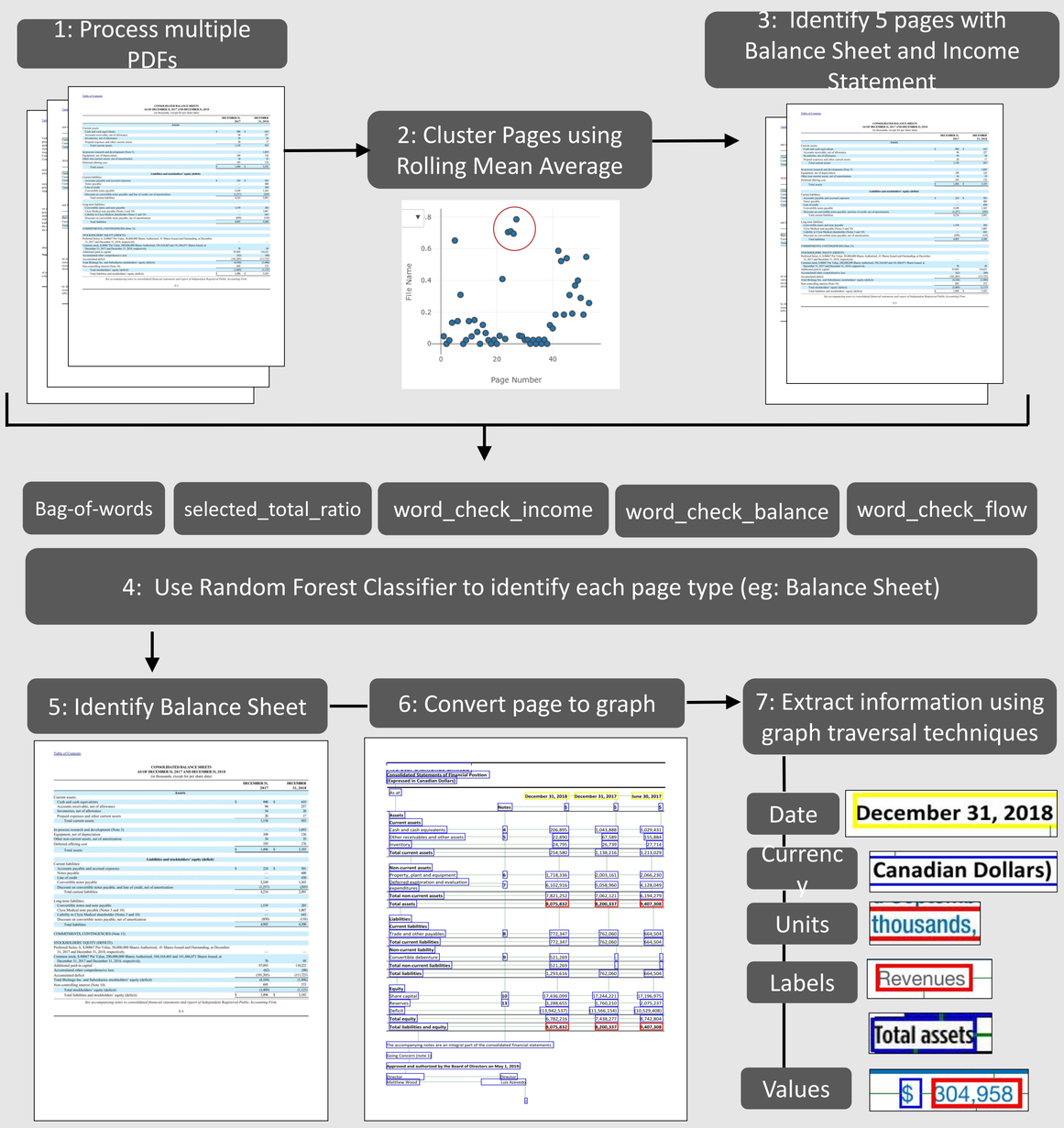
Current Process

- Currently, the analyst manually downloads each financial statement, opens it, identifies the correct statement and extract the value for a financial variable.
- This process is *time consuming, tedious* and has a *high risk of human error.*

Existing Solutions

- The existing methodologies would require us to hardcode each variable location in each page of the pdf to extract required values, which could result in many years of work.

Our Solution



Process

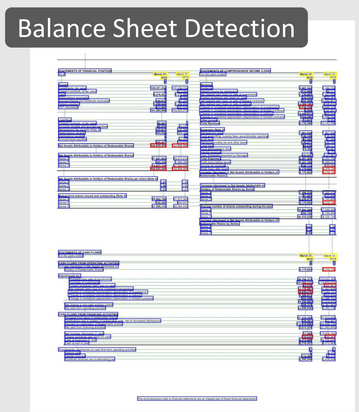
- **Step 1:** Get *Bag-of-words, selected_total_ratio, word_check_income , word_check_balance , word_check_flow flags* for all pages in our sample set.
- **Step 2:** Use *selected_total_ratio* score with rolling mean method to identify the best 5 pages in a group that might contain *Balance sheet* and *Income Statement*.
- **Step 3:** Use the *Random Forest Classifier* model to determine which page is Balance sheet or Income Statement in the group of 5 pages.
- **Step 4:** Convert only those pages to *graph* using a state-of-the-art technique presented by an MIT student at a 2019 NLP Conference.
- **Step 5:** Use graph traversal techniques to extract financial variables.

Evaluation Metrics

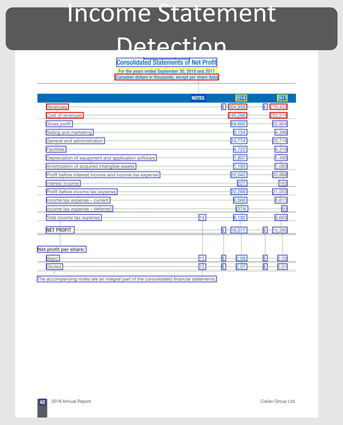
Evaluation Parameter	Sample Size	Positive Count	Accuracy
Balance Sheet Detection	335	324	**96.72 %**
Income Statement Detection	335	323	**96.41 %**
Total Assets Variable Extraction	223	216	**97 %**
Total Revenue Variable Extraction	118	109	**92 %**
Segmentation Page Detection	100	89	**92 %**

Examples

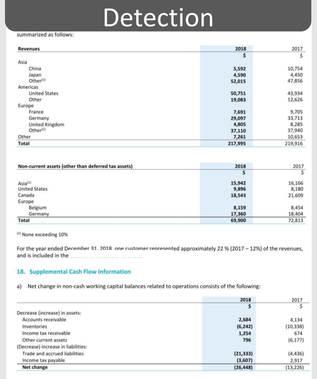
Balance Sheet Detection



Income Statement Detection



Segmentation Page Detection



Web Application Prototype

Benefits and Vision

- *Reduce data redundancy* within the organization by providing a one point solution to access information among different divisions. This would result in saving a lot of storage space and time to download manually.
- *Automate financial variable extraction* process for close to 70000 PDFs per year near real time
- Significantly *reduce the manual hours* spent in identifying and capturing required information by all analysts from these financial statements in Statistics Canada.
- Can be a basis to *develop an interactive web application* that allows analysts across organization to visualize and automatically extract data points

References

- ***GraphIE: A Graph-Based Framework for Information Extraction*** authored by Yujie Qian, Enrico Santus, Zhijing Jin, Jiang Guo, and Regina Barzilay from MIT. They presented this paper at *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies in* June 2019