International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 03 | Mar 2022 www.irjet.net p-ISSN: 2395-0072

Autonomous Platform with AIML Document Intelligence Capabilities to Handle Sensitive Business Information during Merger & Acquisitions

Indranil Dutta

Principal Consultant and Lead Data Scientist

Abstract : *In this research paper, would try to bring some context of Enterprise business practices during Merger and Acquisitions and how AIML can extend capabilities to redact the sensitive and risky business information. This has been a known practice since hundreds of years and lot of organizations are exchanging their secured and sensitive information through different modes to the current owner for the benefit of the business continuity and customer benefits. But all the organizations are imbibed by some internal regulations and protocol whereas the security and maintenance of clients private and sensitive information comes with utmost priority. But die to the process those artifacts need to be carry forwards. Here in this research will share how AIML can help this process with some core benefits through an automated platform to enable a win win situation for both the parties. The paper will talk about some key modules where the AIML capabilities has extensively used to overcome the hindrances.*

Key Words: Sensitive Information, Merger and Acquisition, AIML, business benefit, Information Security and business continuity

1.INTRODUCTION

Now a days all the key cloud service providers have brought powerful APIs with state of art AIML capabilities and most of the organizations are using those APIs to serve their business requirements. But this paper is dedicatedly focus to those small and medium enterprises who are not able to afford the API cost and still need the AIML capabilities to hand the situation with great efficiency. The power of Machine Learning, Cognitive Vision and Natural Language processing all together build a platform capability that can easily handle all the business requirements and regulation maintenance at very effective way. There was always a typical market competition and worry of loosing database to the open-source market agents. Hence considering the customer sensitive as well as enterprise business information are highly private and confidential. As there are majority of the paper works happen during the time of M&A, take overs, Spin outs or Divestment, the AIML capabilities are well handled and smart enough to reduce the manual tasks and probabilities of errors during all the document scans and redactions.

2. BUSINESS CONTEXT

Now its important to understand what the different kinds of information comes under these breaches of sensitive information protocol bracket. Let's discuss the scenario with a context of a Telecom Business operation. Consider A is a small player in the market and X is a market leader. Due to the excessive Infrastructure cost and other factors A decided to divest the Broadband business to X completely. By nature, there would a legal procedure that enables the process of take over from A to X. Now consider A already has a client base of 10 Million people and 60% of then are using their broadband product. Now during the transactions what could be the potential angles business need to look at their database as well as the documents to redact the sensitive information abide by the organizational policy.

Enterprise Angles:

- Product wise revenue margin
- Plan wise revenue margin
- Risk profiling of rolling out the product in market
- Market share
- BTS configuration and IP details
- TCP/IP and PRB package details
- Data quota allocation
- Third party vendor contract terms (Revenue)
- Router switching policy
- Generator configuration to hold maintenance

Customer Angles:

- Customer Name
- Customer contact details (Mobile or Landline no)
- Customer personal of business Email ID
- Customer Passport, Driving license
- Social Security Number, Medicare Number

- Billing details (postpaid connection)
- Recharge details (prepaid connections)
- Data Utilization
- Customer Email Communication records

Now in many a cases same information has used for different context. So, this is also required to understand the context of the problem where it's raised before taking any proactive action.

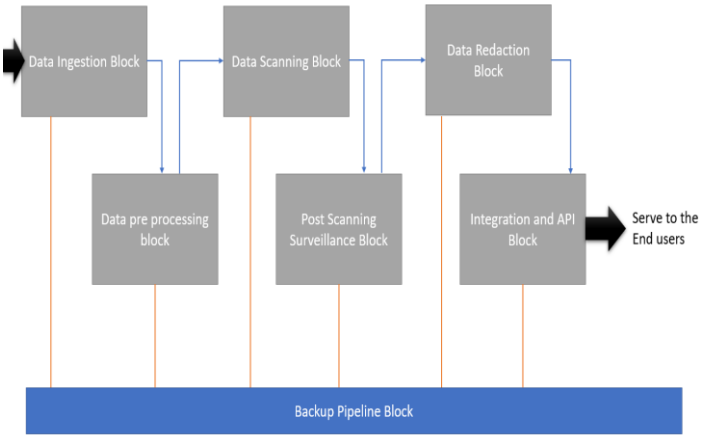
3. PLATFORM as a SERVICES

Here comes the design of the platform that can easily be integrated in any platform whether it's a cloudera/Hortonwork platform or client on premises server or any private/public or hybrid cloud service provider.

There are few key features have considered to design the platform to optimize the performance:

- Performance and accuracy
- Diversification (to cover wide variety of cases)
- Scalability
- Adoptability
- Reusability and flexibility

Figure -1: The Platform architecture high level



There are some core blocks of the platforms.

❖ **Data Ingestion block:**

This is a major block where the capabilities includes capturing structured, semi structured and unstructured data with a variety of more than 36 extension starts from excel, pdf, ppt to HTML, XML, EML to .msg,.eml to JSON, parquets, Avro to images and thumbnails. This is a core block mainly designed for the data engineers who are playing with multiple sources to ingest the data with different formats and keep the API ready so that it has the functionality to plug and play in any environment and using the respective API (ML or non-ML) ingest the data varieties to the core platform and keep in separate bucket for further operations.

❖ **Data preprocessing block:**

The platform is capable of more than 120 operations with those data formats. Some are ML related some are non-ML related. But for different operations different kind of data preprocessing requires. As an example, we are concerned about text extraction of the customer passport images which drives a different data preprocessing task than identifying the key words from a text using Topic modeling or semantic inferencing. More than 435 separate functions have been deployed to cater all 80+ core operations.

❖ **Data scanning block:**

This is the core block for all the major transformations. In this research paper we will be discussing few of the major scanning operations that extensively used state of the art AIML which is embedded as a functionality in the platform. Few of the major modules are:

- Predicting/segregating the signature block from the email body
- Identify the context of using a particular text in a 100 pages document
- Identify the PII data from the customer own documents like passport, Driving license, SSN card, Medicare card.
- Extracting the document layout from a different proforma
- Extracting the attachments from different documents

❖ **Post scanning Surveillance block:**

Lot of the pipelines are executing at a same time and in practical scenario there could be millions of scenarios which the platform has not been experienced earlier which will generate the exceptions logs and failed to process those items. The block is responsible for identifying the test

cases, analyzing the feasibility and pass it on the Preprocessing block so that could be integrated as a module in current platform. That's how the concept is keep on enriching the properties and features based on diversified scenarios. May be 10 years down the line the performance would be optimized and can satisfy near about 100% of the scenarios.

- **Data redaction block:**

  This block actually sits along with the scanning block, the main objective of these block would be applying the redaction to the concerned cases identified by the scanning block and generate the document. AIML process as well as some RPA process are involved in these phases to satisfy the requirement.

- **Integration and API block:**

  The block is mainly responsible for integrating the platform in any new environments and applying the API to consume the data, process and scan it, apply the redaction and restore it back the repository. This is the very complicated block because in practical scenario there could be different environment with different IT configurations, creating the test cases and testing the integration capability is also a major concern. As of now the platform is designed to integrate with Bigdata as well as all major cloud providers. But problems comes when we are trying to integrate the same in customized on-premises server with different architectures. Gradually we will be enriching the platform and integrating the functionalities to adopt the new situations.

- **Backup Pipeline Block:**

  Many a case the API could have failed because of any technical reason. It could be a ML endpoint response or any non-ML end point response. We have identified few scenarios where the possibility of any API failure is. As a backup we will keep a backup pipeline ready. If the API fails, it will trigger the backup pipeline and the functions would get restored. Hence the job continuation chain won't break.

4. PLATFORM CAPABILITY EXPLORATION.

In this section we will discuss some of the major AIML capabilities that the platform provides to handle the task.

a. Predicting/segregating the signature block from the email body

Business Context:

Many a times customers write email to the service desk with all their queries and questions and there has been a chain of communication between them until the issue resolves. Now the customer would have mailed them from his/her personal email id or official email id. For both of the cases there is a chance of capturing the name, mobile no, email id, designation, organization name, organization phone no, address and other important factors. These are denoted as PII and need to be redacted before sharing that information to the new owners.

Old practices:

Based upon certain keywords or keywords matching the signature block is trying to identify. Like – All the context below keywords like "thanks", "thanks and regards", "Best regards" and few other words consider to be the signatures. But lot of time these words are used in the mail body as well and there comes the false positive, many a time sender didn't use any remarks so it's very tough to segregate these signature block s from the mail bodies, even there is a concern that we need to highlight the latest mail block otherwise in every iteration all the old mails will be getting scanned, and duplication happens. Also, there could be potential chance to have some attachments in the mail, we need to download the attachment separately and based upon the document type need to take proactive actions. So, all together this has been a complicated process and just by using some basic keywords and regex matching it's very difficult to identify the entire block with right accuracy.

Potential Solution:

We need a requirement to use the supervised NLP capabilities to meet the requirement. We need to follow the below mentioned steps:

- First need to separate the latest mail conversation from the mail chain. Mailgun talon would be good choice to make that happen without making any further complication.
- Secondly, we need to create the features to support the mail body and signature blocks
- We have created the labelling to form the training data

- Create good data augmentation and add some perturbation to generalize the data
- All the mail body broken up in augmented sentences
- After the labeling its quite clear what are the signature component and what are the body blocks
- Train the model based on the supervised task
- Individual models like CNN1d, GRU, bidirectional LSTM or seq2Seq model won't give promising result
- Need to create a stacked model of all 5 Deep learning algorithm in the training layer, few good ensembles in the meta layer and finally the voting layer to decide the final class.
- This solution really gives a very standard outcome.
- If we increase the training data volume and add diversified perturbations that accuracy of the model would get increased a lot.

Identification:

From the signature block we use some regex logic to identify the Name, Mobile No, city, organization, designation and other related details and redact it and restore the information.

b. Identify the context of identifying comingled keywords from a 100 pages documents

Business Context:

During the time of legal document verification, a bunch of papers with probably 100- or 200-pages documents need to be exchanged which gathers lot of sensitive business information. This is really very hard to guess the context of every keywords and the position as well. But in order to maintain the compliance and the protocol those keywords and sensitive business words needs to be identified and redacted.

Old practices:

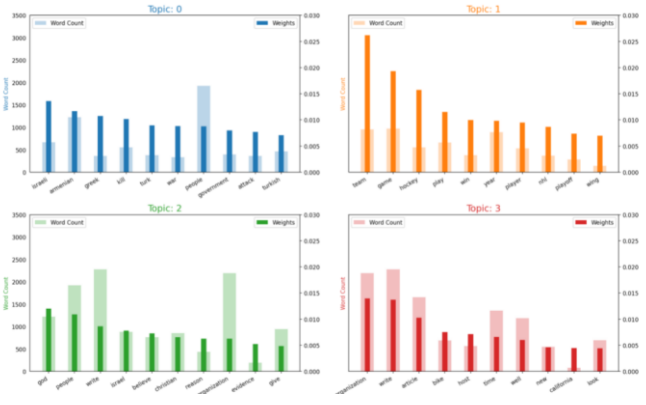
The existing practice was using some basic string matching and regex similarity, wherever there is any matching of the keywords the redaction initiates without justifying the context or the real requirement of hiding the sensitivity of the identified words.

Potential Solution:

There are multiple approaches are there to solve the case but finally the best solution has been taken. The following approaches are:

- Using an abstractive summary to guess the contextual meaning of the particular text section or paragraph. Then from this summarized text use the topic Modeling or semantic inference to identify the most common keywords to guess the context as well as the discuss words. But there is an issue which is losing massive content from the original text because of the abstractive summarization which suppress most of the actual words.
- Another potential solution would be using extractive text summary which uses the sentence similarity matrix and rank and choose top n sentences whose similarity score is high in the matrix. Thereby gives a tentative idea of the entire text but capture the original words from the actual text section. Then apply the topic modeling to identify the keywords.

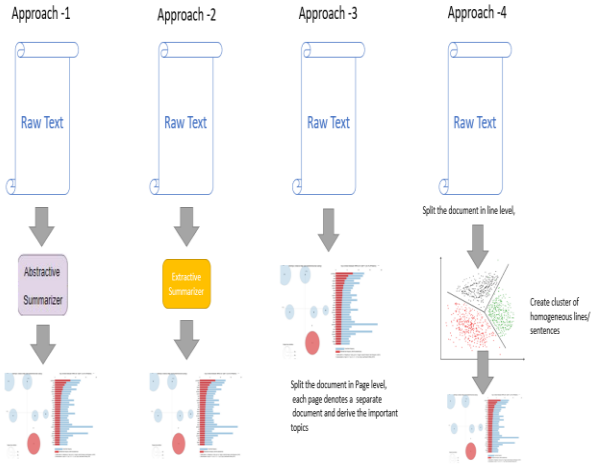
Figure -2: The importance of Topics and its contribution



- Next approach would be document level topic identification. Suppose there are 100 pages, we identify the document definition here is every page, so the topic modeling would identify top 30 keywords that has been discussed in that particular document (i.e. page) and based upon that guess that if any keywords available in this text what could be the potential context for using it and based on that take the decision.

- Other solution could be the same approach as the third one but instead of page as a document we can consider the sentence or even line as a document, thereby we increase the number of documents but get the very granular topics from the respective documents. But this may not be very effective in every context as there are high chances of massive duplications of the same topics in different documents in similar context which actually increase the ambiguity of the overall exercise.

Figure -3: The 4 Approaches of context expose



Identification:

From the extracted topics we need to map the contribution of each topics in respective documents. We will check the topic dominancy in respective documents to guess the potential context and easy to match the keywords in any particular context before initiating the redaction.

Figure -4: The topic dominancy w r t documents

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	3.0	0.9103	organization, write, article, bike, host, time...	[irwin, arnstein, recommendation, summary, wor...
1	1	1.0	0.7182	team, game, hockey, play, win, year, player, n...	[gary, leung, organization, university, toront...
2	2	2.0	0.6801	god, people, write, israel, believe, christian...	[jonathan, hayward, pantheism, organization, w...
3	3	0.0	0.4328	israeli, armenian, greek, kill, turk, war, peo...	[nntp_poste, reply, organization, article, jos...
4	4	3.0	0.5889	organization, write, article, bike, host, time...	[goalie_mask, nntp_poste, ists_ca, organizatio...
5	5	3.0	0.6212	organization, write, article, bike, host, time...	[joe, ehrlich, bmw_moa_member, read, organizat...
6	6	3.0	0.7723	organization, write, article, bike, host, time...	[require, organization, nec_system, article, t...
7	7	3.0	0.7632	organization, write, article, bike, host, time...	[speedy_mercer, look, movie, bike, organizatio...
8	8	2.0	0.9771	god, people, write, israel, believe, christian...	[organization, florida_state, university, foll...
9	9	0.0	0.6571	israeli, armenian, greek, kill, turk, war, peo...	[serdar_argic, day, night, armenian, round, ma...

c. Identify and extract PII data from customer personal documents

Business context:

While making these transactions lot of customers details and required documents needs to be transferred. While opening the connection or accounts he/she has to provide some Id and address proof and the soft copes are embedded inside the repository. But these information are sensitive PII and shouldn't be shared as per the confidentiality and organization policy.

Old practices:

As per the old practice there are some OCR reader which identifies the texts and just redact the texts what ever it would be. But everything shouldn't be the required vase. As an example, a customer belongs to USA and his passport has been issued from Texas which is not a customer related information and should not be redacted. But while making the OCR extraction this information is also been extracted and getting redacted which increase the percentages of False positives.

Potential solution:

As a potential solution we need to follow the below mention steps.

- We created a pool of similar documents and create a MRZ (Machine readable zone) to highlight and annote only that information which are relevant and as per our records.
- Basically it's a semantic segmentation of images as well as object detection considering the part of the image is an object where the relevant PII information persists.
- We have created a diversified training dataset based upon the images and use some preprocessing and augmentation and add some noises to make the training generalized.
- Finally, we have trained the model and compare the results with multiple models and get the best one deployed as a potential solution.

Identification:

The Model will read the correct zone and identify the object with a bounding box and probability score.

Then using a tesseract module convert the object in text and then redact it. There by we are only redacting the concerned information from those secured customer documents.

d. Extract and standardize document layout

This is a non-ML solution and basically concerned about the layout of the documents. There are different types of tables and texts across different types of document. The document layout parser will identify and fit the right layout of the target segments and generate the texts with correct dimensions before making the further preprocessing and redaction.

e. Extract embedded objects attachment inside original documents and treat same chain of operations.

This is also a non-ML solution and very crucial for all types of documents. Many a type of documents like excel csv or outlook files holds embedded object files which could be of different types. Now while processing the main files it needs to be sure that no unstructured file could have hold any embedded object files, surely it has to be a structured file, but the embedded object files could have been different types unstructured images, semi structured raw text or structured table format excel. Now we have to create a pythonic architecture where we create an individual class which will treat the master files and extracts the attachments in a staging folder. Then the child class (using the inheritance property) again calls the same features and process the embedded object files based upon their type. If it's an image then we can directly process as per section c or if it's a word document, excel, pdf or any other outlook file we can use the same logic from master class and process the files to extract the key information embedded inside it.

We have just explored some 5 major modules as the capability of the platform but there are still around 80+ capabilities are there which has already deployed and functional in the platform. Few of them are really very robust AIML solution using state of art techniques but few are very naïve pythonic solution.

5. APPLICATION OF THIS PLATFORM

There is a list of applications of the said architectures:

- M&A for all major players across different markets.
- Property investments for enterprise business
- Business vertical divestment like Future group retail divested to Reliance Retail
- Spin out like Comcast Xfinity home spin out to AT&T
- Divestment like Verizon Aviation to British Airways.
- Joint Venture like Tata group and Walmart for retail e commerce business

Mostly all enterprise engagements these type of application requirement comes into the picture. This platform is a lift and shift in most of all bigdata and cloud platform which is a standard practice for big players. But there is also a flexibility if the platform to be used for small and medium enterprises where the deployment happens on the local server and in their custom IT environments.

6. LIMITATION OF THIS PLATFORM

Gradually once we are exploring in different engagements, we are facing new scenarios which could be potential blocker for these platforms. Let me elaborate few of them.

- There are multiple versions of the MS office file formats. Few of the old versions like 1993-1997 or 1997-2003 doesn't supports some functionalities which is a declared and common drawback of the Microsoft office licenses. We are using the office latest version, hence this could be a potential limitation.
- Some of the small and medium enterprises they are using their own server where some of the open-source packages are not executed, hence the model gets stuck. This is not a major issue while playing the same functionality in cloud or bigdata environment but do in custom on premise servers, this is another potential limitation of the platforms.
- As this framework is designed keeping in mind all the groups of clients, hence there is no features like Kubernetes has considered. Now for very big players like HSBC Holding and UBS transactions a huge chunk of load placed on the platform which significantly concern the performance issue.

This are some key limitations of the current release, gradually the features are taken into consideration while declaring the subsequent releases.

7. BENEFITS OF THIS PLATFORM

- A collaborative feature to address the entire business problem through a platform as a service rather than accumulating bits and pieces solutions.

- Adoptable for all scale of organization for their own usage.
- Flexible in Azure, GCP, AWS, IBM cloud, Oracle Cloud, SAP cloud, Big data environments.
- Easy maintenance and bug fixing and easy tracking of the exceptions through proper log management.
- Less expensive than directly use cloud APIs.
- Support 80+ functionality to cover almost all functionalities involve to the related domain.
- Cover 11 Line of business Keywords while creating the bag of words for all NLP solution including banking, insurance, telecom, manufacturing, retail, automobile. Logistics and few others.

8. CONCLUSION

Now a days it's a common practice to develop lot of sophisticated edge products based on state of art AIML solutions but the ultimate ROI was never justified. This platform is basically concerned about more than $200billion annual transactions over the globe per year across 150 countries. Every year large players combinedly has to pay near about $40 billion as a penalty for non-maintenance of organizational protocols and local government rules. Now it gives a very potential vibes about the responsibility of the platform. This platform alone with all its capability can save 15%-20% of the annual fees that organization has to bear because of non-maintenance and fraudulent disclosure and breech of confidentiality which is almost equal to a global revenue of a CMM Level-5 organization. In subsequent release the loop wholes have been integrated tightly to make this as a complete end to end solution and a global success in the market of Merger and Acquisition.

REFERENCES

1. Redaction Of Documents – Substance & Its Application - Intellectual Property - India (mondaq.com)
2. Document Redaction: What It Is and How It Works | Record Nations
3. Redacting Documents and Records.pdf (dataprotection.ie)
4. Data exposure: Using software to redact personal data from public documents | Computerworld
5. Redact | Practical Law (thomsonreuters.com)

BIOGRAPHIES

Indranil Dutta is a Principal consultant and Lead Data scientist with more than 11 years of rich experience in Data science and Artificial Intelligence in various industries. His core competency is in delivering scalable Data science and AI projects in Big data and Cloud environments.