Loan Repayment Prediction using Big Data

Executive Summary

Loan Repayment Survey has become a trending topic of discussion in a plethora of banks due to the significantly rising loan fraud cases. For example, over the last few years, many Australian banks have been hit by several putative mortgage frauds because of the increment in the cost of buying a house. Therefore, a lot of people are implementing illegal activities just to secure a mortgage, as a consequence, most of the banks are facing losses because of their traditional & inefficient methods of Loan Repayment Prediction.

So, the key idea is to extract the essential knowledge about loan applicants from vastly growing Big Data (scattered across various systems) that would skyrocket the accuracy of Loan Repayment Prediction which could be backed up with justifiable private and open data. More than 70% of Big Data includes text, images, audios, videos, documents, etc. (Unstructured & Semi-Structured data) available in various Data Islands as per the type of data required i.e. Social media, Internet of Things (IoT), Public or Private. The major challenge in analyzing Social & IoT data is the transformation of initial raw data captured from several data islands into curated form, i.e., contextualized data and knowledge that is maintained and made available for use by end-users and applications. Data curation requires identifying relevant data sources, extracting data and knowledge, cleaning, maintaining, merging, enriching and linking data and knowledge. To tackle this challenge, I introduce Knowledge Lake (a contextualized Data Lake), it automatically curates raw social & IoT data for generating business insights. In this paper, I presented a reliable combination of CrowdCorrect & iProcess data curation pipelines, to enable analysts to engage with both social & IoT data efficiently for uncovering hidden patterns and generating insights. On one hand, CrowdCorrect offers automatic data curation, feature extraction, crowdsourcing, Domain Knowledge, and Linking services for Social data. On the other hand, iProcess allows analysts to ingest data from IoT enabled devices, retrieve knowledge from this data and link it to process (execution) data. In this paper, I analyzed the social behavior of a loan applicant from social media & IoT-enabled devices, to highlight how the pipelines significantly improves the quality of extracted knowledge. This will undoubtedly help the banks in storytelling about the personality of a loan applicant.

Contents

List of Figures

1. Industry context – Domain

In Banking, a loan means property, money, or other material products given to another gathering in return for future reimbursement of the advance worth or principal amount, alongside interest or account charges. That implies the bank just makes benefit (premium) if the borrower returns the loan amount. In case a loan applicant doesn't reimburse the money, the moneylender faces financial loss. Loans are usually issued by financial institutions, money related establishments, and governments. There are two types of loans i.e. secured or unsecured. Mortgages are the best example of a secured one because it is backed by collateral. Unsecured credits commonly have higher financing costs than secured ones, as they are less secure for the moneylender. With a secured loan, the loan specialist can repossess the security on account of default. The two most basic inquiries in the loaning business are i) How dangerous is the borrower? ii) Should we loan him/her? [1] [2]
There are several data islands of Big Data i.e. social, open, private, IoT, etc. In this research paper, I used Social and IoT data islands due to their incredible assistance in optimal decision-making by building a clear image of the social personality of a loan applicant. Twitter (https://twitter.com/home), Facebook (https://www.facebook.com/), LinkedIn (https://www.linkedin.com/feed/), & Instagram (https://www.instagram.com/) are the major components of Social Data Island, however, CCTVs, sensors, etc are for IoT Data Island.

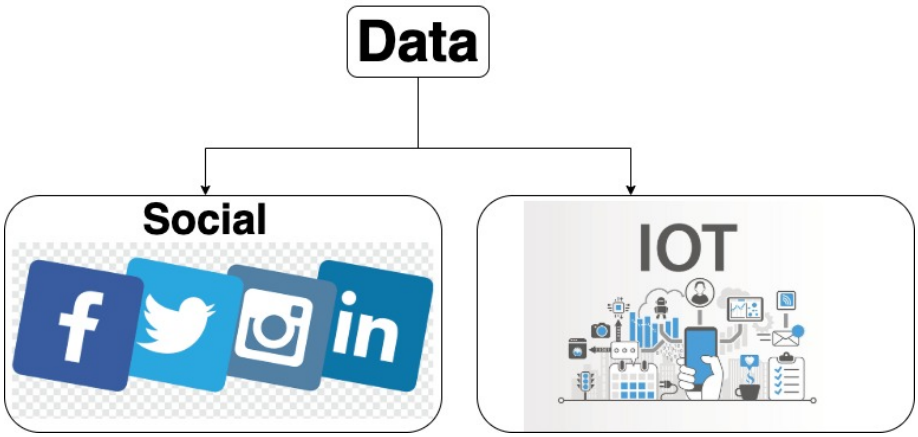

Fig. 1. Essential Data Islands for analysing the behaviour of a Loan Applicant

Social data means the information shared by social media users online, which incorporates metadata, for example, the user's area, language, biographical data, as well as shared links. Social data is vital to advertisers searching for customer bits of knowledge that may build deals or, on account of a political battle, win votes. There are numerous kinds of social information, including tweets from Twitter (https://twitter.com/home), posts on Facebook (https://www.facebook.com/), pins on Pinterest (https://www.pinterest.co.uk/?autologin=true), posts on Tumblr (https://www.tumblr.com/), etc. [3]

IoT is the network of physical devices/objects augmented with Internet-enabled computing devices to enable those objects to sense the real-world, has the potential to transform many industries. [4] The definition of the Internet of things has evolved due to the convergence of multiple technologies, real-time analytics, machine learning, commodity sensors, and embedded systems. [5]

2. Motivating Scenario

Despite recent Loan Repayment Prediction Techniques i.e. applying Machine Learning on applicant's financial details (Structured data), verifying Credit History, etc., cases of 'liar loans' in Australia are still surging drastically, especially secured loans. A liar loan is an advance that has been given to a borrower in the wake of giving false information about the borrower's compensation, property, and work or business history to a Bank. Such a strategy infers various people who may not, regardless, be certified for an advance get one. [6]
Up to 33% of Australian home loans could be "liar loans" in view of verifiably wrong data, investment bank UBS had cautioned. The worldwide financial monster has followed up a survey of home loan borrowers in 2017 when it discovered proof of widespread mortgage fraud. Just 67% of home loan applications were "totally truthful and exact". Around 40% of loan applications through a home loan intermediary had at least one deception. Most of the NSW residents lied on their loan application. [7]

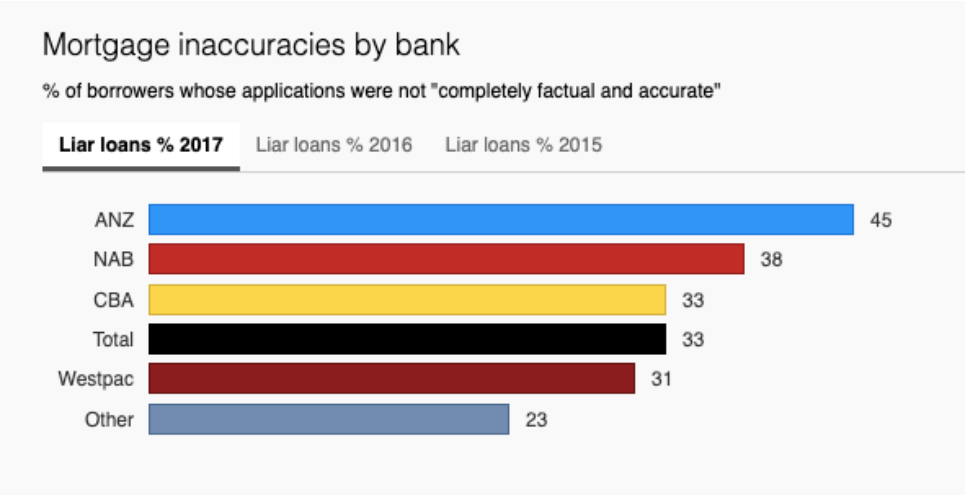

Fig. 3. Mortgage inaccuracies in four major Australian banks

In Fig 3, each of the four significant banks recorded a lot more significant level of dodgy mortgage loan endorsements than the remainder of the financial sector (at 23%). Consequently, because of the critical increment in loan fraud cases, we should set our concentration towards hugely expanding Big Data also, with the goal that we can make sense of the social behavior and other fundamental IoT based information of the loan applicant. The questions that we can answer based on Big Data are:-

- What about the social behavior (sentiment) of the applicant over sometime regarding loan?
- What about the personality graph of a loan applicant?
- How often the loan applicant visit his/her property?

3. Problem statement

Banks provide loans to borrowers in exchange for the promise of repayment with interest. That means the banks only make a profit if the borrower pays off the loan. However, if he/she doesn't repay the loan, then the bank will face loss. This type of situation comes under loan fraud cases.

In spite of using several top-notch Machine Learning algorithms (KNN, Lin Reg, SVM, etc.) for predicting the loan repayment, the fraudulent cases are still rising dramatically. The fundamental explanation for this could be the unrecognized social conduct or inappropriate property assessment of the loan candidate. For instance, in 2017 UBS led a study for recognizing the surmised extent of 'liar loans' cases. ANZ Bank had the most noteworthy extent of home loans that were not "totally genuine and precise", with 45 percent falling into the investment bank's class of "liar loans", a term borrowed from the US subprime mortgage crisis. That was fundamentally over the business average of 33 percent, similar to NAB's degree of loan endorsements that contained deceptions or inaccuracies.[6] That implies, rather than concentrating on simply organized information we should likewise infer some viable bits of knowledge utilizing Unstructured information about the loan applicant. In this way, we can expand the precision of loan reimbursement prediction procedures to considerably diminish the forthcoming financial loss by 'liar loans' cases.

3.1 Current (problem) state

The present loan reimbursement forecast process concentrates just on the finance-related data (Structured) of the candidate to predict if the borrower will reimburse the advance by its maturity date or not by utilizing a lot of Machine Learning Algorithms. Notwithstanding, there is no authoritative guide of which algorithms to utilize given any circumstance. What may take a shot at certain data sets may not really work on others. Therefore, consistently assess methods utilizing cross-validation to get solid estimates. Sometimes we might be eager to give up some improvement to the model if that would build the unpredictability significantly more than the percentage change in the improvement to the assessment metrics.[2]

Hence, instead of just playing with numbers we would shift our focus towards drastically growing Big Data which would surely help in creating intellectual personality graphs of our loan applicants for uncovering the hidden stories rather than just working on numerical data.

3.2 Desired (goal) state

Big Data has several data islands i.e. social, open, private, IoT, etc., but here we use only Social and IoT data islands.

3.2.1 Social Data Island

Fig. 4. Social data curation techniques

Using some scalable algorithms we transform social items (e.g., a Tweet in Twitter) into semantic items, i.e., contextualized and curated items. In our approach to social data curation, we specifically focus on cleansing and correcting the raw social data of loan applicants; and present a pipeline to apply curation algorithms (automatic curation) to the information items in social networks and then leverage the knowledge of the crowd as well as domain experts to clean and correct the raw social data. [8] Then, we will perform data extraction, Sentiment Analysis, & Knowledge Graphing to derive some meaningful insights.

Therefore, after analyzing the social data of a loan applicant we can plot a graph as shown in Fig. 5, depicting his/her behavior (sentiments) regarding loaning over a specific time period.

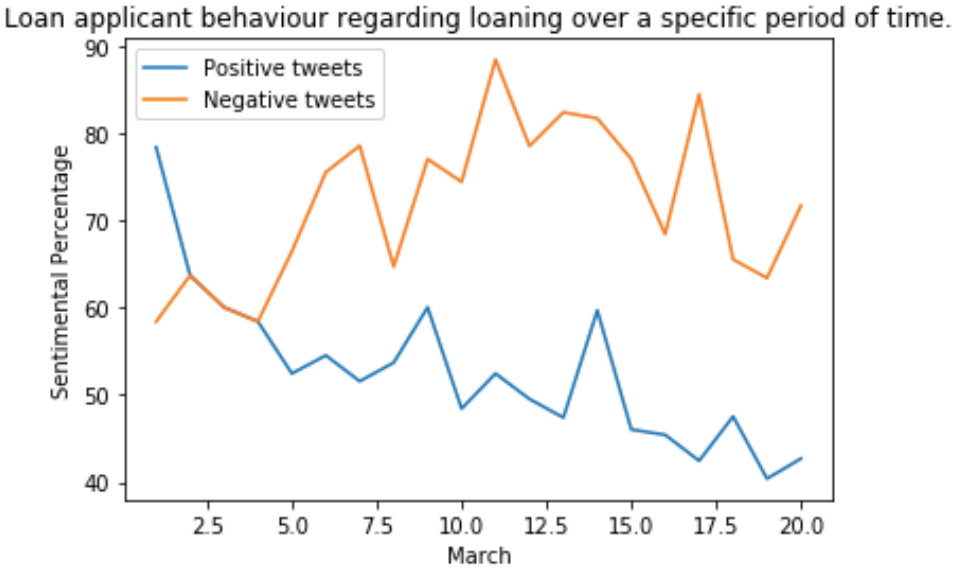

Fig. 5. Social behaviour of a loan applicant

3.2.2 IoT Data Island

Fig. 6. IoT- enabled devices

Empowering IoT information in business process investigation, as introduced in this paper, a novel way to deal with upgrade information-driven procedures for improving risk-based dynamics in knowledge-intensive procedures. The tale thoughts of Knowledge Lake and narrative, introduced in this paper, empower us to put the initial move towards empowering storytelling with process information. This will enable analysts to ingest data from IoT devices(drones, CCTVs, and more), extract knowledge from this data and related the data to process analysis. Summarization techniques is a novel approach to enable analysts to understand and relate the big IoT and process data to process analysis in order to communicate analysis findings and supporting evidence in an easy way. The proposed approach will enhance data-driven techniques for improving risk-based decision making in knowledge-intensive processes.

Initially, we will store the photograph & property related details of the loan applicant in a location-enabled Drone. So, the drone would be sent to exact location and fetch data from all the CCTVs within a range of 5 km of a particular time period. This entails gathering information about the person including physical appearance, and activities in the physical environment of the person, person's activity data such as phone calls and emails, and information on the person detected by sensors (e.g. CCTVs).

Therefore, after analyzing the IoT enabled- devices (CCTVs, sensors, etc) near the property of a loan applicant we can plot a graph as shown in Fig. 7, depicting how often he/she visit over a certain period of time.

Fig. 7. Presence of a loan applicant near his/her property

4. Approach

To address this challenge, CoreDB (https://www.mongodb.com/atlas/data-lake): a Data Lake as a Service, to recognize data sources (IoT, Private, Social and Open) and ingest the big process data in the Data Lake. CoreDB works with numerous database management systems (relational to NoSQL), offers an implicit plan for security and following, and gives a solitary REST API to sort out, record and question the information and metadata in the Data Lake. Here, we use a combination of CrowdCorrect and iProcess data analytics pipelines to deal with both Social and IoT data efficiently as shown in Fig. 8.

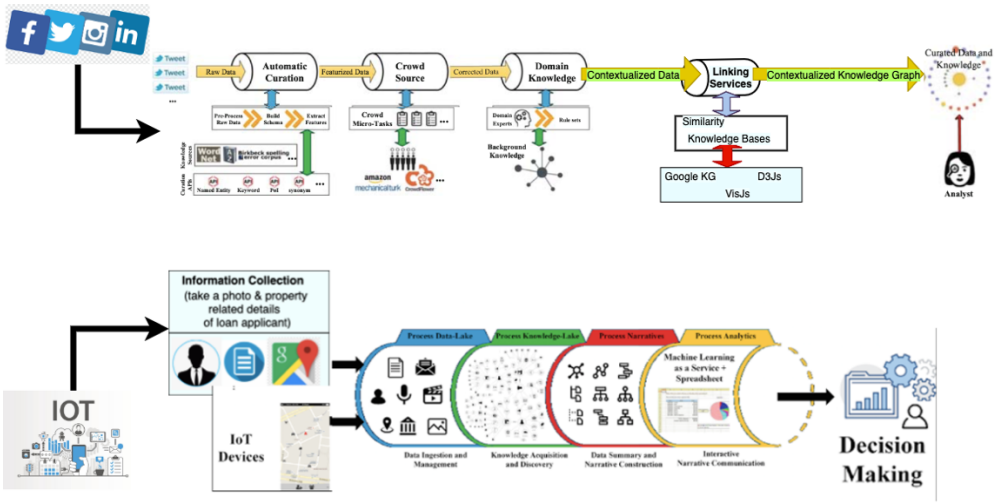

Fig. 8. A blend of CrowdCorrect & iProcess pipelines

4.1 CrowdCorrect (Social Data)

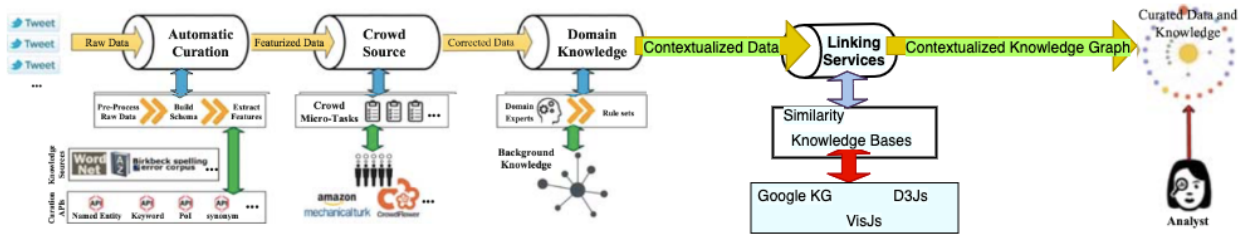

Fig. 9. Curation pipeline for cleaning, correcting & graphing social data

To comprehend the social information and supporting the dynamic procedure, it is essential to address and change crude social information created on interpersonal organizations into contextualized information and information that is kept up and made accessible for use by examiners and applications. To accomplish this objective, we present an information curation pipeline, CrowdCorrect, to empower investigators to purge and curating (planning) social information for dependable business information examination. Figure 9 delineates an outline of the CrowdCorrect curation pipeline, comprise of three principle information preparing components: Automatic Curation, Crowd Correction, Domain Knowledge Reuse, and Linking Services.[8]

4.1.1 Automatic Curation: Cleansing and Correction Tasks

Data cleaning manages to recognize and expelling errors and inconsistencies from data so as to improve the nature of data [9]. In social networking, this job is more challenging as social users generally use abbreviations, acronyms, and slangs that can't be recognized utilizing learning algorithms. At this progression, we structure and execute three services: data ingestion, extraction, and correction.

4.1.1.1 Data Ingestion Service

For this service, we use Apache Nifi (https://nifi.apache.org/) to obtain streaming social data (live tweets) of our loan applicant from Twitter for immediate use and storage in a database as shown in Fig. 10. In Apache Nifi, we create the data flow by simply drag and drop interface without doing any programming & our data will go through processors where each processor transforms the data. Nifi is scalable across a cluster of machines to increase the throughput & also provide guaranteed delivery so that we'll not lose any data.

So, in this processor group i.e. Twitter to MongoDB Atlas, I used are GetTwitter (to get the specific tweets as per the country, loan-related hashtags, the applicant used id, & language of our loan applicant), EvaluateJsonPath (to get only required attributes), UpdateAttribute (to change the names of attributes), ReplaceText (to replace the content), MergeContent (to merge all the .json tweet files), PutMongo (to store the transformed data in MongoDB) processors for ingesting & cleaning the raw data.

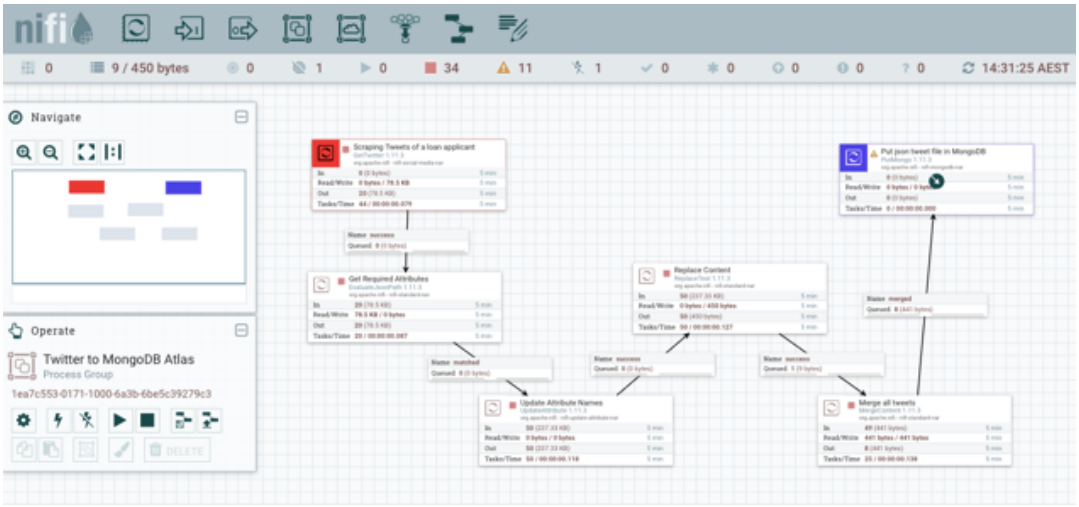

Fig. 10. Ingesting Tweets using Apache Nifi

As a result, a JSON (json.org/) file containing all the desired tweets of our loan applicant i.e. Malcolm Turnbull (https://en.wikipedia.org/wiki/Malcolm_Turnbull) as per his loaning hashtags, user id, language, & location will be stored in a MongoDB Atlas's Data Lake due to the ease of organizing, indexing and querying the data and metadata as shown in Fig. 11.


```
{
"Tweet Id":"570828097240432640",
"Text":"Thrilled to receive a copy of the Bank of NSW Board minute recording a loan to ancestor John Turnbull in 1817 https://t.co/2MgBAB37lv",
"Name":"Malcolm Turnbull",
"Screen Name":"TurnbullMalcolm",
"Created At":"Thu Feb 26 06:09:58 +0000 2015",
"Favorites":"36",
"Retweets":"12",
"Language":"English",
"Client":"<a href=\"http://twitter.com/download/iphone\" rel=\"nofollow\">Twitter for iPhone</a>",
"Tweet Type":"Tweet",
"Media Type":"",
"Media URLs":"",
"URLs":"1",
"Hashtags":"3",
"Mentions":"2"
}
```


Fig. 11. Fetched tweets file in json format

For example, according to this Twitter schema, a tweet in Twitter have some crucial attributes such as (i) Text: The text of a tweet; (ii) geo: The location from which a tweet was sent; (iii) hashtags: A list of hashtags mentioned in a tweet; (iv) Language: The language a tweet was written in, as identified by Twitter; (v) links: A list of links mentioned in a tweet; (vi) Media Type: The type of media included a tweet; (vii) mentions: A list of Twitter usernames mentioned in a tweet. [8]

4.1.1.2 Data Extraction Service

The curation APIs empower developers to effortlessly include features - for example, extracting part of speech, keyword, and named entities such as Organizations, Persons, Locations, Companies, Products, Diseases, etc as shown in Fig. 12. These features incorporate, but not restricted to:

- Lexical features: words or jargon of a language, for example, Keyword, Topic, Phrase, Abbreviation, Special Characters (for example '#' in a tweet), Slangs, Informal Language and Spelling Errors.
- Natural-Language highlights: elements that can be separated by the investigation and combination of Natural Language (NL) and discourse, for example, Part-Of-Speech (for example Action word, Noun, Adjective, Adverb, and so forth.), Named Entity Type (for example Individual, Organization, Product, and so forth.), and Named Entity (i.e., an occurrence of an element type, for example, 'Malcolm Turnbull' as a case of element type Person).
- Time and Location include the notices of time and area in the substance of web-based life posts. For instance, on Twitter, the content of a tweet may contain a period notice '3 May 2017' or an area notice 'Sydney; a city in Australia'. [8]

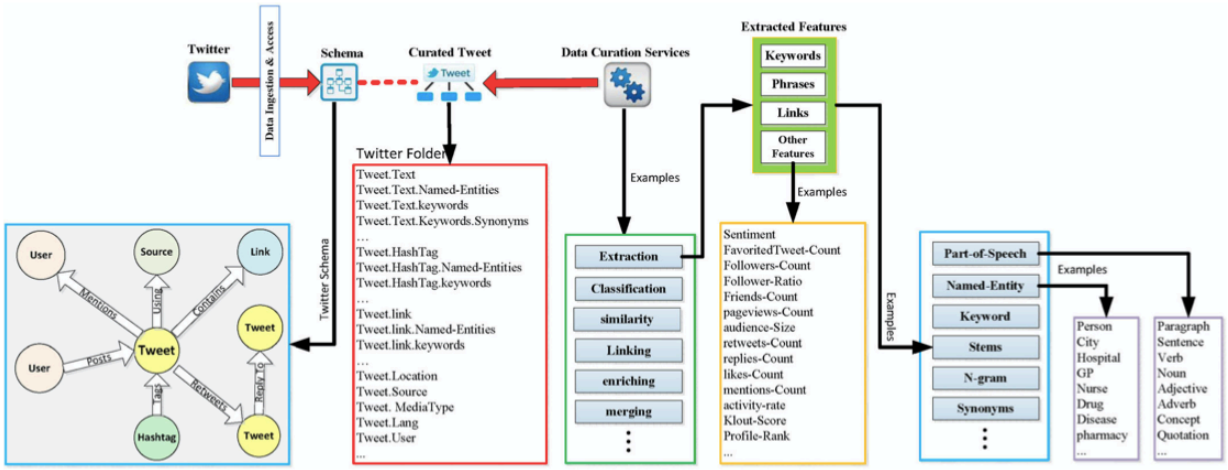

Fig. 12. Feature extraction on Twitter

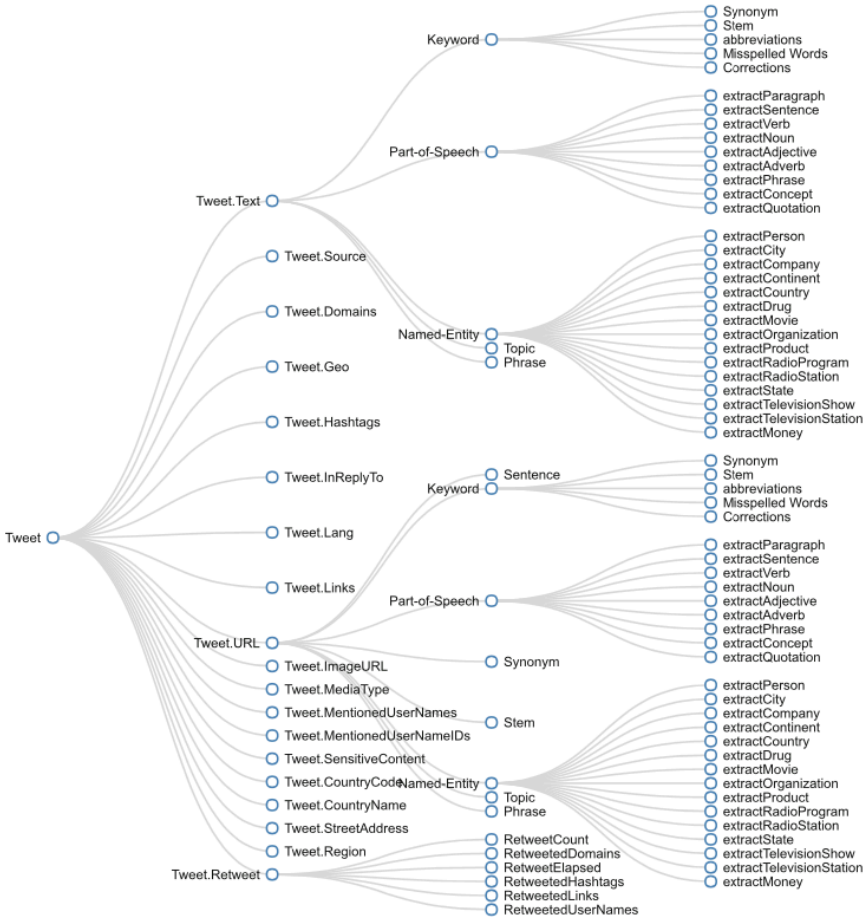

Fig. 2. An example of an automatically curated tweet.

4.1.1.3 Data Correction Services.

To recognize and address the incorrect spelling, jargon (for example extraordinary words or articulations utilized by a calling or gathering that are hard for others to comprehend) and shortened forms. These administrations influence information sources and administrations, for example, WordNet (wordnet.princeton.edu/), STANDS4 (abbreviations.com/abbr api.php) administration to recognize abbreviations and shortened forms, Microsoft subjective services to check the spelling and stems, and cortical (cortical.io/) administration to distinguish languages. The aftereffect of this progression (programmed curation) will be a comment on a dataset that contains the cleaned and adjusted crude information. [8]

4.1.2 Crowd Sourcing

Social items, for example, a tweet on Twitter, are generally written in structures not fitting in with the standards of grammar or accepted utilization. In like manner, social items become text standardization challenges as far as choosing the best possible strategies to identify and change over them into the most precise English sentences.

[10] Crowdsourcing [11] strategies can be utilized to get the information of the crowd as a contribution to the curation task and to tune the programmed curation stage.

Crowd Correction. Crowdsourcing quickly assembles huge quantities of individuals to achieve undertakings on a worldwide scale. For instance, anybody with access to the Internet can perform micro-tasks [12] (little, measured assignments otherwise called Human Intelligence Tasks) on the request for seconds utilizing stages, for example, Amazon's Mechanical Turk (mturk.com), crowdflower (crowdflower.com/). It is additionally conceivable to utilize social administrations, for example, Twitter Polls or basically structuring a Web-based interface to impart the micro-tasks to companions and associates. We have structured two sorts of crowd micro-tasks: suggestion and correction tasks. [12]

4.1.2.1 Suggestion Micro-tasks We design and implement an algorithm to present a tweet along with an extracted feature (e.g. a keyword extracted using the extraction services in Sect.4.1.1) to the crowd and ask the crowd worker if the extracted feature can be considered as misspelled, abbreviation, or jargon.

```
Data: Automatically Curated Social-Item
Result: Annotated Social-Item
Extract Features from Social-Item;
array Question-Set ["misspelled ?", "abbreviations ?", "jargon ?"];
for Each feature in Extracted-Feature do
    for Each question in Question-Set do
        Generate Suggestion Micro-Task as follows:
        Display Social-Item;
        Display feature;
        Display question;
        Retrieve the Crowd Feedback (Yes/No);
        Annotate the feature (e.g. "misspelled" or "not misspelled");
    end
end
```

Algorithm 1. Automatically generating Suggestion Micro-Tasks.

4.1.2.2 Correction Micro-tasks. Subsequent to knowing the crowd decision about the extracted feature can be considered as incorrectly spelled, shortening, or jargon; request the right type of the component. The crowd will be approached to choose the right suggestion or information another remedy if necessary. Algorithm 2 delineates how we naturally produce correction micro-tasks.

4.1.3 Domain Knowledge

In the last stage, there could be some cases where the crowd would not be able to correct the features. So, there would be some cases where the meaning of a keyword or an abbreviation is domain-specific. For example, in the tweet "They gave me the option of AKA or limb salvage. I chose the latter.", the automatic and crowd correction tasks can identify AKA as an abbreviation (AKA stands for 'Above-knee amputation'), however providing a correct replacement for this term requires the

domain knowledge in health.

```
Data: Annotated Social-Item (Suggestion Micro-Tasks)
Result: Corrected Social-Item
Extract Features and Annotations from Annotated Social-Item;
for Each feature in Extracted-Feature do
    for Each annotation in Annotation-Set do
        if annotation = ("misspelled" OR "abbreviations" OR "jargon") then
            Generate Correction Micro-Task as follows:
            Display Social-Item;
            Display feature;
            Correction-Set = Correction-Service(feature);
            Display Correction-Set; Display Question("Choose/Input the
            correct" + annotation);
        else
            Annotate the Social-Item("No Need for Manually Correction");
        end
    end
end
```

Algorithm 2. Automatically generating Correction Micro-Tasks.

To address this test, we offer a domain-model interceded technique to utilize the knowledge of domain specialists to distinguish and address things that could not be revised in previous steps.

4.1.4 Linking Services

4.1.4.1 Similarity

Data matching usually relies on the use of a similarity function f(v1,v2) → s, which can be used to assign a score s to a pair of data values v1 and v2. These values are considered to be representing the same real-world object (extracted keywords, named entities, etc) if s is greater than a given threshold t. So, analysts may need a collection of similarity APIs to measure the Cosine similarity of two vectors of inner product space and compare the angle between them, the Jaccard similarity of two sets of character sequence, the length of the longest common subsequence between two strings using an edit distance algorithm, the hamming distance between two strings of equal length and more.

4.1.4.2 Sentiments

It is the procedure of computationally distinguishing and classifying conclusions communicated in a bit of text, particularly so as to decide if the client's mentality towards loaning is positive, negative, or neutral.[13] There are a few APIs accessible that can without much of a stretch calculate the appropriate sentiment of a given amount of text as shown in Fig. 13.

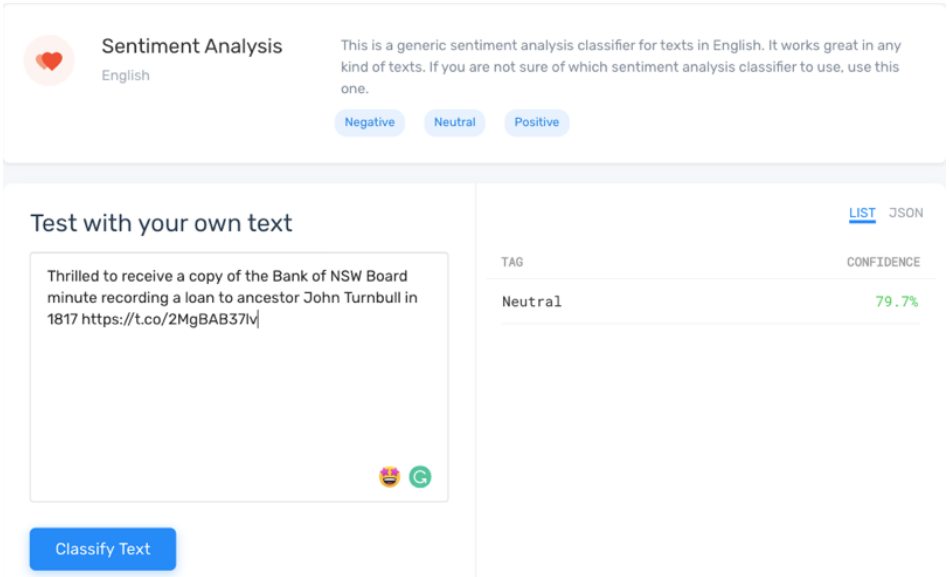

Fig. 13. An example of sentiment analysis using an API service

4.1.4.3 Knowledge Bases

While extracting various features (e.g. named entities, keywords, synonyms, and stems) from the text, it is important to go one step further and link the extracted information items into the entities in the existing Knowledge Graphs (e.g. Google KG and Wikidata). For example, consider that we have extracted 'M. Turnbull' from a tweet text. It is possible to identify a similar entity (e.g. 'Malcolm Turnbull') in the Wikidata (https://www.wikidata.org/wiki/Wikidata:Main_Page). As discussed earlier, the similarity API supports several functions such as Jaro, Soundex, QGram, Jaccard and more. To achieve this, we have leveraged the Google KG (https://developers.google.com/knowledge-graph) and Wikidata APIs to link the extracted entities from the text to the concepts and entities in these knowledge bases as shown in Fig. 14. For example, the Google API (https://developers.google.com/apis-explorer) call will return a JSON (https://en.wikipedia.org/wiki/JSON) file that may contain the URL.

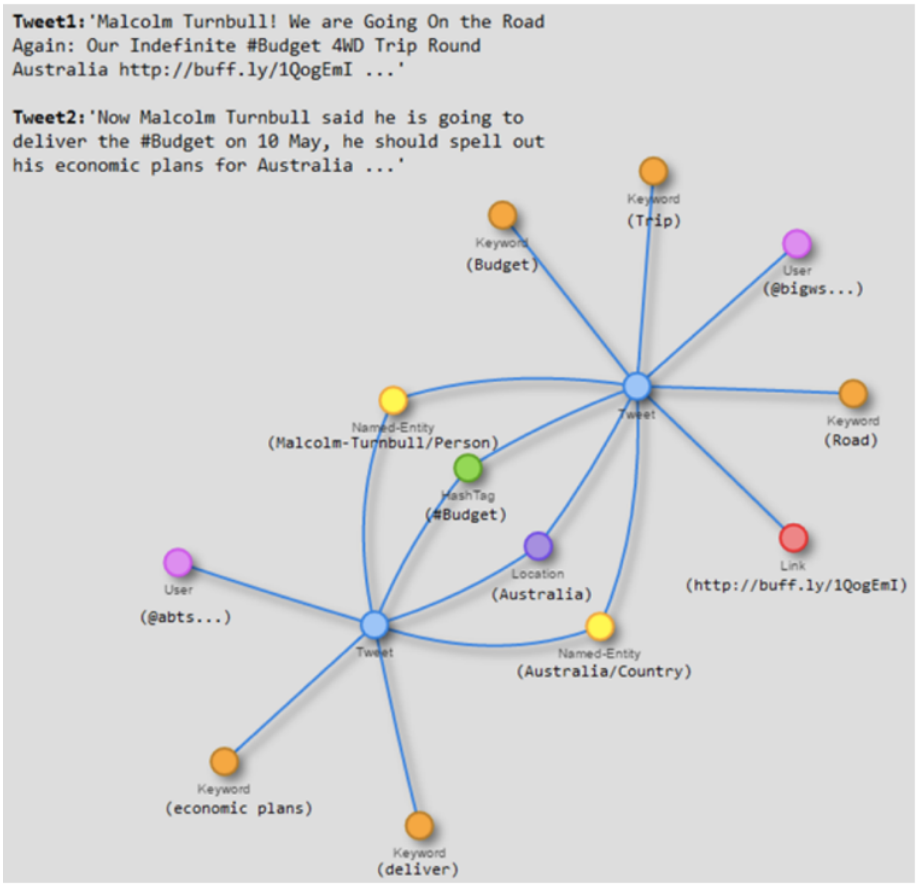

Fig. 14. Use extracted features from Twitter to link related tweets

4.2 iProcess (IoT Data)

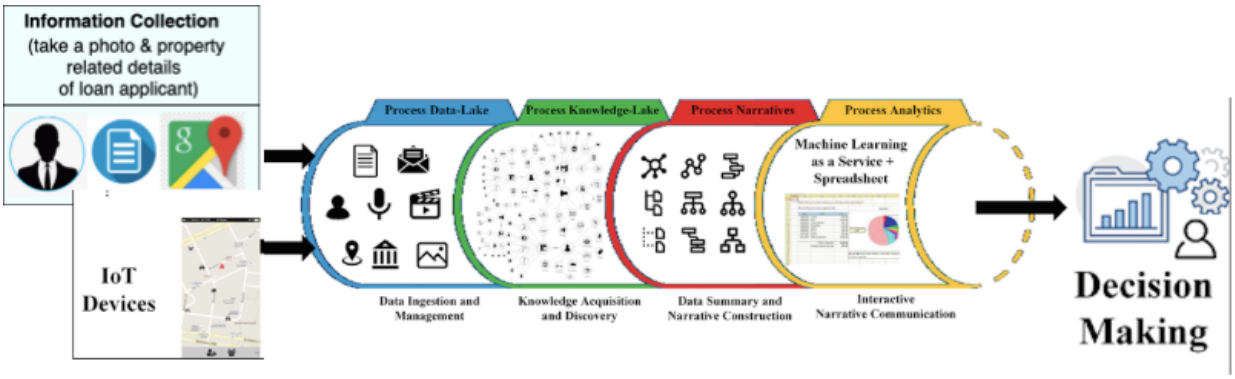

Fig. 15. iProcess pipeline

4.2.1 Process Data-Lake

Here, we use Apache Nifi (https://nifi.apache.org/) Flow to ingest & process IoT devices (location-enabled drone & CCTVs) data as shown in Fig. 16. Initially, we will store the photograph & property related details of the loan applicant in a Drone. So, the drone could reach the exact location and fetch the data from CCTVs within a

range of 5 km of a particular time period. Hence, we can find all the cars and people in these CCTVs near the respective property.

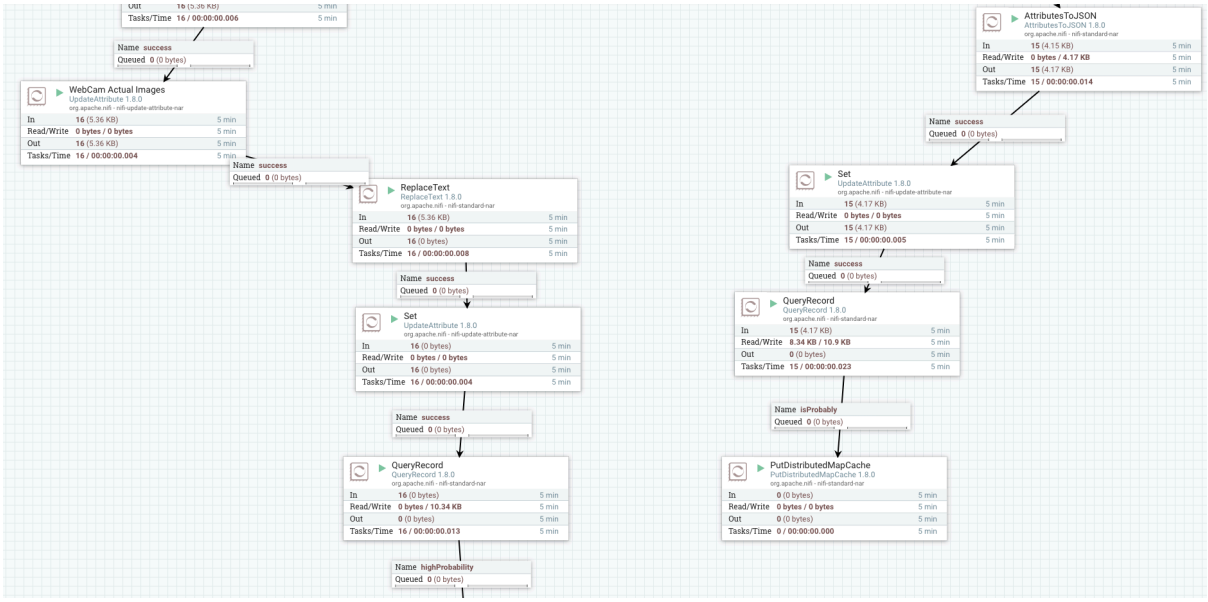

Fig. 16. Ingesting CCTVs data using Apache Nifi

As a result, we will get a JSON file containing a list of all the URLs from nearby CCTVs as shown in Fig. 17.

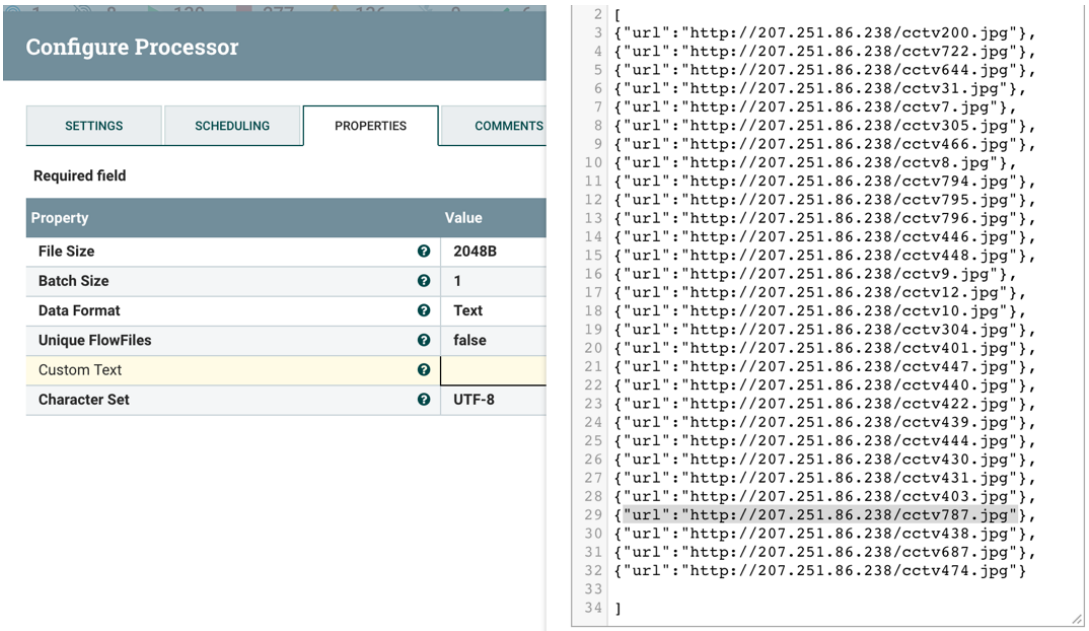

Fig. 17. Ingested .jpg format images from CCTVs data

Finally, we will transmit this data to our respective Data Lake in MongoDB Atlas which manages multiple database technologies (from relational to NoSQL), offers a built-in design for security and tracking, and provides a single REST API to organize, index and query the data and metadata in the Data Lake.

4.2.2 Process Knowledge-Lake

Data Lake (https://www.mongodb.com/atlas/data-lake) stores crude data and let the data analyst conclude how to curate them later. We present the thought of Knowledge Lake [14], i.e., a contextualized Data Lake, to give the establishment for big data analytics via consequently curating the crude data in the Data Lake and to set them up for determining insights. The Data Curation APIs [15] in Knowledge Lake give curation tasks, for example, extraction, linking, summarization, annotation, enrichment, classification and more; classifying, indexing, sorting and categorizing data - into the data and knowledge continued in the Knowledge Lake. [16]

This will enable us, for example, to extract and link information about the loan applicant's IoT data sources and to relate them. The goal of this phase is to contextualize the Data Lake and turn it into a Process Knowledge- Lake which contains: (i) a set of facts, information, and insights extracted from the raw data; (ii) process event data i.e., observed behavior; and (iii) process models, e.g., manually or automatically discovered. To achieve this goal, we present a graph model to define the entities (process data, instances, and models) and the relationships among them. [16]

4.2.3 Process Narratives

In this phase, we present an OLAP [17] style process data summarization technique as an alternative for querying and analyzing data. Here, the system will be able to use interactive (artifacts, actors, events, tasks, time, location, etc.) summary generation to select and sequence narratives dynamically. This novel summarization method will enable process analysts to choose one or more dimensions (i.e., attributes and relationships), based on their specific goal, and interact with small and informative summaries. This will enable the process analysts to analyze the process from various dimensions. Figure 18 (B) shows a sample OLAP dimension.

In OLAP [17], cubes are defined as a set of partitions, organized to provide a multi-dimensional and multi-level view, where segments considered as the unit of granularity. In our situation, process cubes can empower viable examination of the Process Knowledge Graph from alternate points of view and with numerous granularities. For example, by aggregating and relating all evidence from the loan applicant and his/her property.

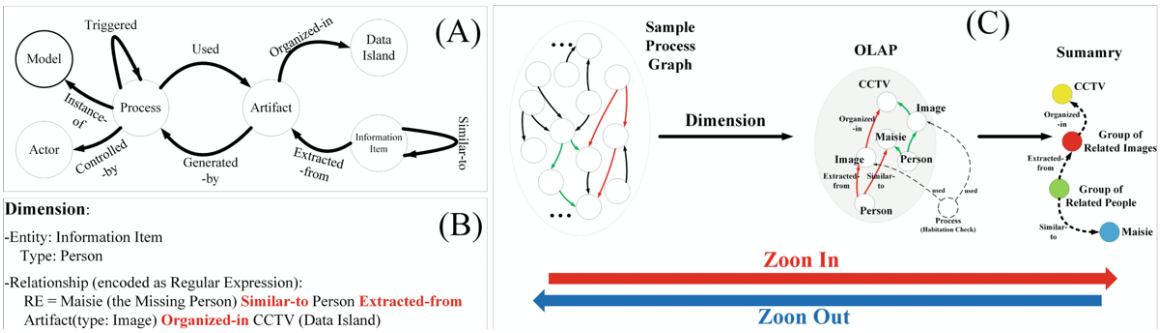

Fig. 18. Process Knowledge Graph schema (A), a sample OLAP Dimension (B) and an interactive graph summary (C).

The formalism of the summary S will empower to think about various dimensions and perspectives of a narrative, including the occasion structure (narratives are tied in with something occurring), the reason for a narrative (narratives actors and artifacts), and the job of the audience (narratives are abstract and rely upon the viewpoint of the procedure analyst). Likewise, it thinks about the significance of time and provenance as narratives may have various implications after some time. We build up an adaptable outline age calculation and bolster three kinds of summaries. Figure 19 shows the versatile summary generation process. Tailing we present these summaries:

- **Entity Summaries:** We use relationship conditions to summarize the Process Knowledge Graph-dependent on the arrangement of dimensions originating from the attributes of node entities. Algorithm 1 in Fig. 19, will produce all conceivable entity summaries. For instance, one potential outline may incorporate every related picture caught in a similar area from CCTVs. Another summary may include every related picture caught in the equivalent timestamp.
- **Relationship Summaries:** We use correlation conditions to summarize the Process Knowledge Graph-based on a set of dimensions coming from the attributes of attributed edges. Algorithm 2 in Fig.19, will generate all possible relationship summaries. For example, one possible summary may include all related relationships typed controlled-by and have the following attributes "Controlled-by (role = 'Investigator'; time = 'τ1'; location = '255.255.255.0')".
- **Path Summaries:** We use path conditions to summarize the Process Knowledge Graph-based on set of dimensions coming from the attributes of nodes and edges in a path, where a path is a transitive relationship between two entities showing a sequence of edges from the start entity to the end. Algorithm 3 in Fig. 19, will generate all possible path summaries. For example, one possible relationship summary includes all related images captured in the same location and contains the same information item, e.g., the loan applicant.

4.2.4 Process Analytics

In this phase, we make a set of machine learning algorithms available as a service on top of the scalable summary generation framework to enable the analysts to manipulate and use the summaries in spreadsheets to support these three operations: (i) roll-up: to aggregate summaries by moving up along one or more dimensions, and to provide a smaller summary with less details. (ii) drill-down: to disaggregate summaries by moving down dimensions; and to provide a larger summary with more details; For instance, in Fig. 20, applying the drill-down procedure on CCTV1 (information source) will give a progressively definite summary, gathering all the things over various focuses in time. (iii) slice-and- dice: to perform selection and projection on depictions. For instance, to see CCTV depictions originating from 2 dimensions for example time and location. The slice-and-dice operation can be essentially observed as a regular expression that bunches together unique entity and relationship summaries (introduced in the spreadsheet tabs) and weaves them together to build way summaries, outlined in Fig. 20

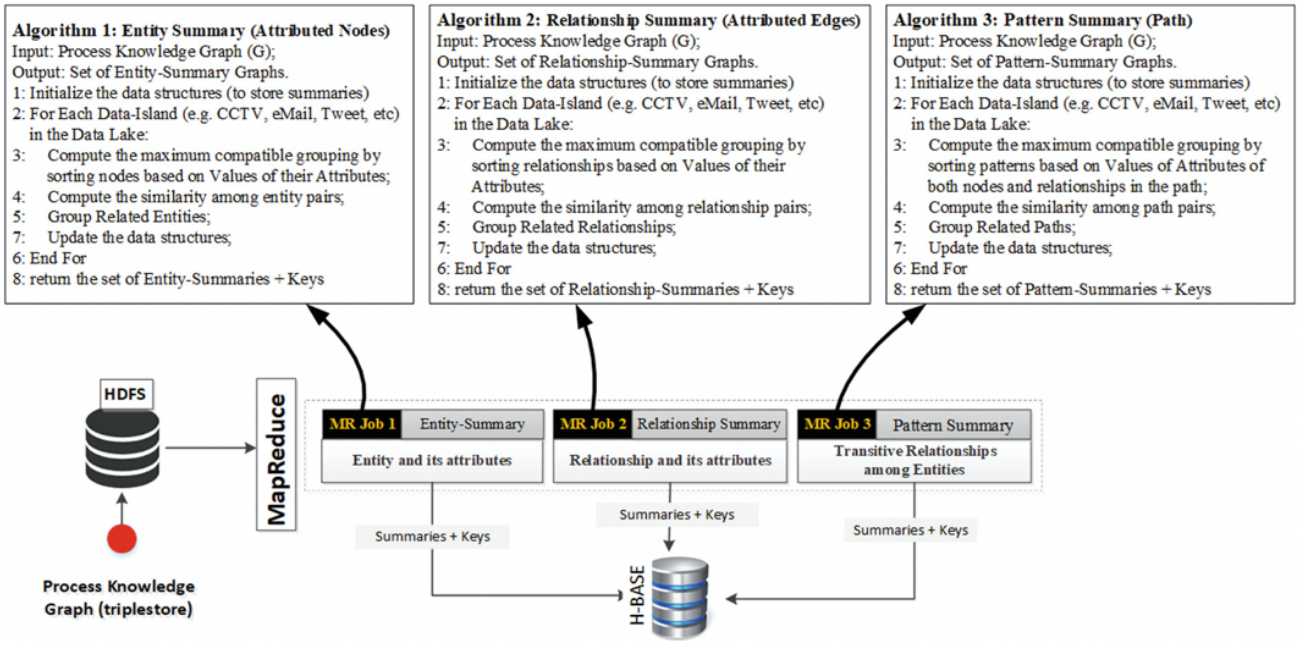

Fig. 19 Scalable Summary Generation

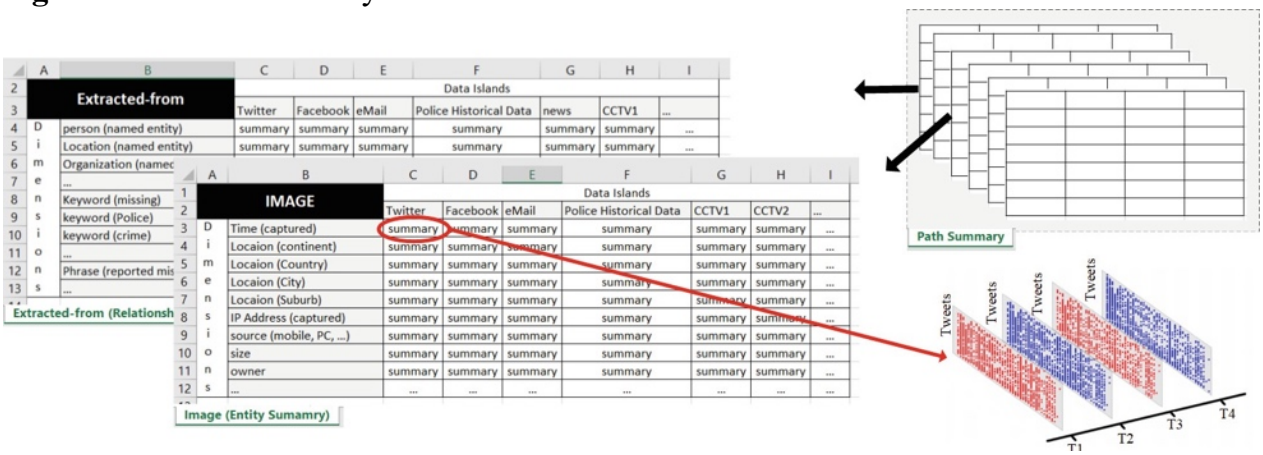

Fig. 20 Presenting a spreadsheet like interface on top of the scalable summary generation framework.

5 Results

After evaluating the Social & IoT data of the loan applicant, we can derive plenty of meaningful insights about him/her which will undoubtedly assist Business Analysts in making optimal decisions. For example, if our loan applicant is Mr. Malcom Turnbull then after analyzing all his Social & IoT data (relevant to the loaning process) we will get the following overview of his personality as shown in Fig. 21.

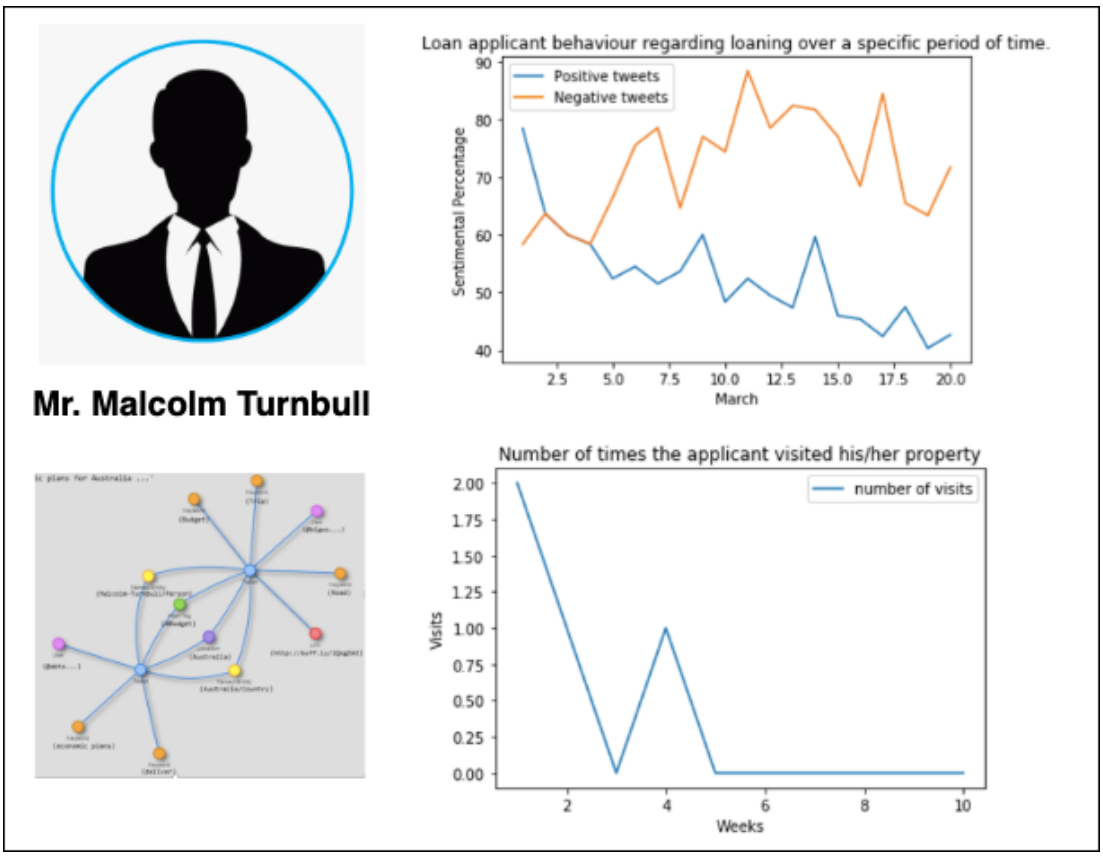

Fig. 21. Overview of a loan applicant personality

Here, we will get to know about the following:-

• Social behavior of a loan applicant over a specific period of time-related to loaning activities. As we can observe here, the positive social behavior of our applicant is decreasing significantly throughout the month of March. That means he could be not interested in repaying the borrowed loan after some time.

• Named Entities, Verbs or Keywords linked to the applicant financially.

• How often the applicant visited his property over a specific period of time. In the graph shown above, we can clearly see that applicant did not visit his property even once in the last 5 weeks.

6 Conclusion

These days, Loan fraud cases are surging drastically due to the inefficient and traditional loan repayment prediction methods. The main reason behind this could be the unnoticeable social behavior of a loan applicant over a particular timestamp. So, our key focus would be on extracting crucial knowledge from remarkably raising Big Data that would make a significant surge in the accuracy of Loan Repayment Prediction. Approximately, 70% of the Big Data includes Unstructured data (audio, text, video, etc) which is present in various data islands e.g. Social, IoT, Public,

Private, etc. In this paper, we focused on Social & IoT data islands because they provide a large amount of user-generated data on a continuous basis. The process of transforming raw data into contextualized data (curated data) is not an easy procedure as it includes figuring out the vital data sources, extracting data and knowledge, cleaning, maintaining, merging, enriching and linking data and knowledge. While curating the social data we introduced a Data Analytics Pipeline i.e. CrowdCorrect, which is a blend of crowdsourcing and algorithmic methods to deal with redundancies, abbreviations, slangs, misspellings, etc. Nevertheless, for IoT data we used iProcess, for data ingestion, knowledge extraction, linking, data summarizing using OLAP, & using Machine Learning as a service.

We presented a scenario in understanding the social behavior of our loan applicant by analyzing his Social & IoT based data. We evaluated our approach by analyzing his/her social behavior, physical presence near the property, & connections with named entities related to the loan. Therefore, we can present a personality overview of a loan applicant. To improve the loan applicant personality dashboard, as future work, we could work on a novel Platform-as-a-Service that makes it simple for developers & analysts of all expertise levels to utilize AI innovation, the manner in which individuals utilize a spreadsheet.

References

1. https://www.investopedia.com/terms/l/loan.asp
2. https://towardsdatascience.com/predicting-loan-repayment-5df4e0023e92
3. https://www.investopedia.com/terms/s/social-data.asp
4. Beheshti A. et al. (2018) iProcess: Enabling IoT Platforms in Data-Driven Knowledge-Intensive Processes. In: Weske M., Montali M., Weber I., vom Brocke J. (eds) Business Process Management Forum. BPM 2018. Lecture Notes in Business Information Processing, vol 329. Springer, Cham
5. https://en.wikipedia.org/wiki/Internet_of_things#cite_note-Linux_Things-1
6. https://www.smh.com.au/business/banking-and-finance/nab-stung-by-alleged-liar-loan-fraud-20181004-p507si.html
7. https://www.abc.net.au/news/2017-09-11/500b-dollars-of-liar-loans-in-australia-ubs/8892030
8. Beheshti, A., Vaghani, K., Benatallah, B., & Tabebordbar, A. (2018). CrowdCorrect: A Curation Pipeline for Social Data Cleansing and Curation. Information Systems in the Big Data Era, 24–38. doi:10.1007/978-3-319-92901-9_3
9. Rahm, E., Do, H.H.: Data cleaning: problems and current approaches. IEEE Data Eng. Bull. 23(4), 3–13 (2000)
10. Sosamphan, P., et al.: SNET: a statistical normalisation method for Twitter. Mas- ter's thesis (2016)
11. Howe, J.: The rise of crowdsourcing. Wired Mag. 14(6), 1–4 (2006)
12. Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., Horton, J.: The future of crowd work. In: CSCW (2013)
13. https://www.google.com/search?sxsrf=ALeKk01DIVjsBvGd2cIxdfQX-QiG8KgMPg%3A1586243475605&ei=kyeMXq6yJMSl9QPE3YPwBw&q=sentiment+analysis&oq=sentiment+analysis&gs_lcp=CgZwc3ktYWIQAzIECCMQJzIECCMQJzIECAAQQzIECAAQQzIECAAQQzIECAAQQzIECAAQQzICCAAyBAgAEEMyAggAOgQIABBHSg0IFxIJMTEtMTc5ZzExSgoIGBIGMTEtMWcyUNslWNslYOAqaABwAngAgAGtAYgBrQGSAQMwLjGYAQCgAQGqAQdnd3Mtd2l6&sclient=psy-ab&ved=0ahUKEwjux4-34dXoAhXEUn0KHcTuAH4Q4dUDCAw&uact=5
14. Beheshti, A., Benatallah, B., Nouri, R., Tabebordbar, A.: CoreKG: a knowledge lake service. In: Proceedings of the VLDB Endowment (PVLDB 2018), vol. 11(12) (2018). https://doi.org/10.14778/3229863.3236230
15. Beheshti, S., Tabebordbar, A., Benatallah, B., Nouri, R.: On automating basic data curation tasks. In: WWW (2017)
16. Beheshti, A., Schiliro, F., Ghodratnama, S., Amouzgar, F., Benatallah, B., Yang, J., … Motahari-Nezhad, H. R. (2018). iProcess: Enabling IoT Platforms in Data-Driven Knowledge-Intensive Processes. Business Process Management Forum, 108–126. doi:10.1007/978-3-319-98651-7_7]
17. Beheshti, S., Benatallah, B., Motahari-Nezhad, H.R.: Scalable graph-based OLAP analytics over process execution data. Distrib. Parallel Databases 34(3), 379–423 (2016)