Whitepaper On

GENERATING CREDIT OPINION WITH SEMANTIC TECHNOLOGY AND SUPERVISED LEARNING

Joint Efforts Of

Mizuho Bank, Ltd. (Singapore) & Latize Pte Ltd (Singapore)

November 2020

Table of Content

Executive Summary

- Effective Credit Risk Management (CRM) is the backbone of any commercial bank's risk management framework. The credit assessment process requires the assessor to derive an opinion on the credit worthiness of a borrower using a diverse range of information such as exposures, financial performance, company developments and external business environment. The information gathering process is often manual and time-consuming.

- Our POC aims not only to help streamline the data information gathering process, but also relook at the way data sources can be organized and visualized in a more insightful manner using knowledge graph and semantic ontology.

- With a credit grading recommendation engine and an early warning flag built within the system using advanced data driven methods like machine learning, we aim to generate an educated credit opinion using deep learning models (i.e. LSTM and SEQ2SEQ). Ultimately, credit risk evaluation would evolve into a process that is more time sensitive and predictive.

- The positive outcome of our PoC suggests that the introduction of an engine equipped with basic cognitive ability similar to a human being would be a milestone towards in the development of explainable AI. Not only will the engine produce a predictive outcome for human consumption, it will also be able to explain how it managed to derive at the conclusion.

1. Conceptual Overview of the Project

1.1. Re-engineering Credit Risk Management Workflow:

Effective Credit Risk Management (CRM) is the backbone of a commercial bank's corporate counterparty risk management framework. In Singapore, banks are required to conduct regular and systematic reviews of all credit facilities (including off balance-sheet items) that it has extended to its borrowers, adhering minimally to the prudential standards set out in MAS Notice 612. After taking into consideration all relevant internal and external information pertaining to the borrower, each facility will be assigned a MAS Grading (i.e. Pass. Special Mention, Substandard, Doubtful, Loss)[1] with an accompanying credit opinion.

Credit risk monitoring process comprises several activities to stay ahead of the any unseen credit developments. These activities include but not limited to Close Watch Exercises and Early Warning Reporting to keep taps with the latest developments. Typically, an analyst would constantly search through various available source of systems to analyse and update relevant information (e.g. Client's financial statements, industry/business updates, latest company news peer comparisons etc.), before making a recommendation on the Grading.

Handling large volume of credit risk-related information in the age of information explosion can be a daunting experience even for highly-skilled credit analysts, who typically plough through various information sources to keep up with changes in a borrower's creditworthiness in order to make insightful credit evaluation.

This project seeks to re-engineer and streamline a typically manual processes by introducing new technologies to augment the credit assessment of bank's borrowers from assigning a MAS Grading to generating a credit opinion.

[1] MAS Notice 612 Credit Files, Grading and Provisioning (Last Revised Date: 29 December 2017)

1.2. **Credit Risk Monitoring in an Age of Information Explosion**

Typically, credit analysts would have to process a massive amount of structured and unstructured data, to develop an opinion after careful examination of the financial data. **S**ome of the types of information an analyst needs to ingest while performing credit risk assessment activities include

Macroeconomic Trends	Data and information that allow one to measure and track the performance, structure, and behavior of the entire economy. This could pertain to significant economic, environmental, or geopolitical event that can influences a regional or national economy.
Market Trends	High frequency data and information pertaining to changes in interest rates, currencies, equity, bond and commodity prices, Market movement of these indicators are highly volatile.
Company News	Company news reported or broadcast through traditional and new media sources. The volume and the diversity of the information flowing through social media would still require verification to ascertain the authenticity of reports.
Expert Opinions	Brokers' reports, market analysis and external rating agency's guidance and commentaries.

Where the process is mostly manual it becomes time-consuming and can give rise to data quality issues and affect the timeliness of the credit risk monitoring process. Systems storing structured data, unstructured data, multiple data formats, non-standardized financial statements and inconsistent text formats of various financial information lead also exacerbate the challenge.

1.3. **Benefits of AI- Powered Solution:**

The current manual process can be intelligently automated by introducing an expert system to address all the mentioned challenges. The proposed system is to sharpen credit risk assessment quality by covering all data sets possible, and speed up the turnaround time in report generation without compromising required standards of counterparty credit risk management.

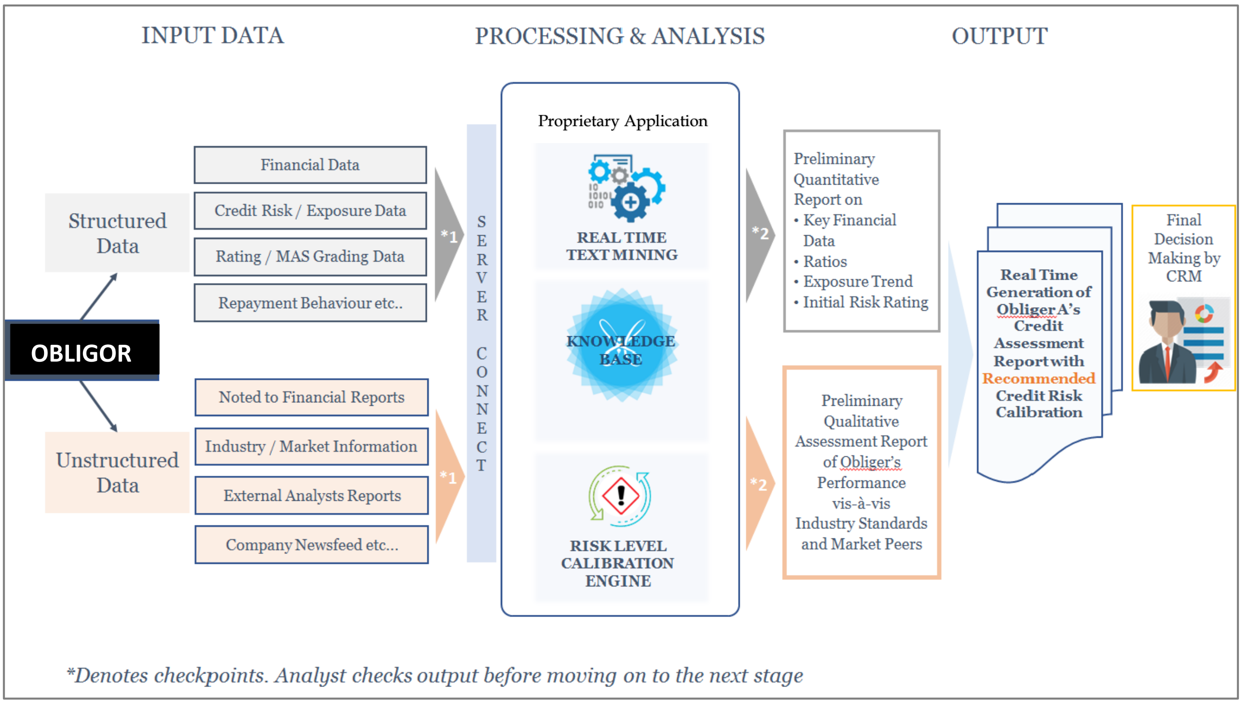
The solution would be a platform that allows for seamless connectivity and accessibility amongst several data sources used for day-to-day credit assessment and monitoring. It would be an AI-enhanced engine that is capable of assigning a MAS Grade to each borrower, flag out early-warning signs of potential financial weakness based on and generate a credit opinion consistent with the Grading recommendation.

The PoC used a semantic[2] technology platform to enhance NLP/NLG techniques to improve the accuracy of the outcomes. Using semantic technology-based NLP and NLG the system adopts a more deterministic approach that relies on ontology for discovery, disambiguation and information sequencing and reinforces these models through appropriate Machine Learning algorithms (probabilistic) with limited training document sets.

This is the unique value proposition and the use of a semantic technology engine that combines deterministic (knowledge base) and probabilistic (machine learning) methods to derive business outcome. Figure 1 is a graphical summary of the conceptual framework of the new process flow:

[2] Semantic platform refers to data being stored in a form that enables systems to understand data. Semantic technologies provide an abstraction layer above existing IT technologies that enables bridging and interconnection of data, content, and processes (https://en.wikipedia.org/wiki/Semantic_technology).

Figure 1 : A New Conceptual Framework of CRM Workflow



The core advantages of this new framework are as follows:

Timeliness: A complete data pipeline to ingest, extract, process and transform the data as or when required;

Robust and Effective solution to the complex problem: A system that is able to comprehend and digest the information from the Knowledge Base built on the Data Aggregation Platform;

Traceable and validation-ready: An explainable AI outcome that constructs the preliminary credit recommendation and present the rationales is consistent with the risk calibration. Over time with sufficient available data, improve the provided rating recommendation;

Early Warning System: The ability to process the huge amount of data to generate early warning flags for dynamic credit risk monitoring;

2. Design of PoC Technical Platform:

In today's fast changing environment, organizations need timely and actionable insights from their data but are often constrained by a fragmented enterprise information landscape where data exists in silos – disparate and isolated.

While many unanticipated questions and information requirements arise, users are disconnected from the data they need for decision making. This leads to a focus on ad hoc reports and time consuming, expensive IT integration.

To resolve these data issues, the PoC uses Latize's Ulysses as core platform to build the solution. Ulysses is an 'agile decision engine' which combines artificial intelligence and semantic technology to automatically link related entities such as persons, companies, financials, products, etc. across multiple systems and sources.

2.1. Semantic Network

A semantic network[3] is a graph notation for representing knowledge in patterns of interconnected nodes. Semantic networks became popular in Artificial Intelligence (AI) and Natural Language Techniques (NLP) because it has the ability to represent knowledge and support reasoning.

A Typical Semantic Network Chart

Knowledge can be stored in the form of graphs, with nodes representing objects in the world, and arcs representing relationships between those objects. Semantic nets, which consist of nodes, links and link labels, are referred to as associative nets as the nodes are associated with other nodes.

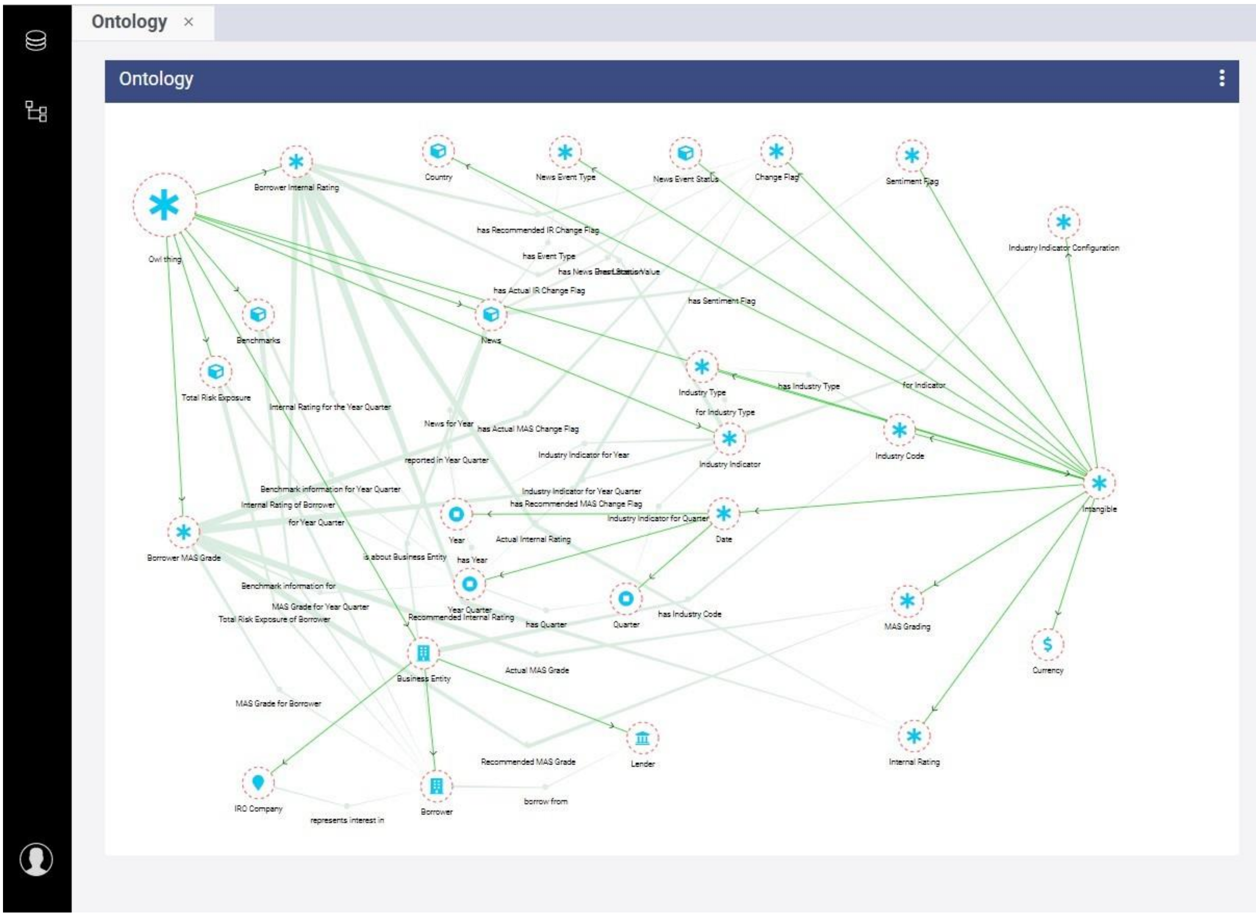
[3] https://www.sciencedirect.com/topics/computer-science/semantic-network

2.2. **Knowledge Graph**

A knowledge graph is a programmatic way to model a knowledge domain with the help of subject-matter experts, data interlinking, and machine learning algorithms. Knowledge Graphs allow companies to connect the dots in their business and see the big picture for evaluating risk and value.

Built as large semantic networks, Knowledge Graphs enable companies to semantically integrate diverse data and draw connections at an unprecedented scale. They also allow users to connect external sources of data efficiently, regardless of their underlying data formats and models. (Figure 2)

Figure 2: A Graphical Representation of a Knowledge Graph in Ulysses



They enrich the context of data by incorporating domain-specific knowledge vocabularies, taxonomies or ontologies. Ulysses is expressive enough to even extract hidden relations for accurate text classification, annotation, entity resolution and relationship disambiguation. The effective design and implementation of Knowledge Graphs requires 3 components:

- Bridging diverse data silos regardless of data formats, serializations, conceptualizations, and technology ecosystems;
- Investigating interconnected data to find out insightful patterns; and
- Deriving context-relevant knowledge from the large amounts of integrated data.

Adopting the Ulysses platform helps in exploring the hidden relationships among the data points for better understanding to the user. This deep exploration is quite fast and computationally inexpensive. It's easy to explore two hop or more down to establish the relationship.

2.3. **Building the Engines and Modules**

While the focus of this PoC is on Credit Opinion Module, it is important to provide a brief explanation of the Grading Recommendation Engine to provide a clearer context to the end-to-end data pipeline and workflow preparation.

In order to work on generating a credit opinion, the starting point is (i) a Grading Recommendation Engine that ingests relevant company and financial information to generate a Grading recommendation. The output of this engine will then go through to (ii) a Credit Opinion Module that construct an opinion based on the recommendation.

A) Grading Recommendation Engine:

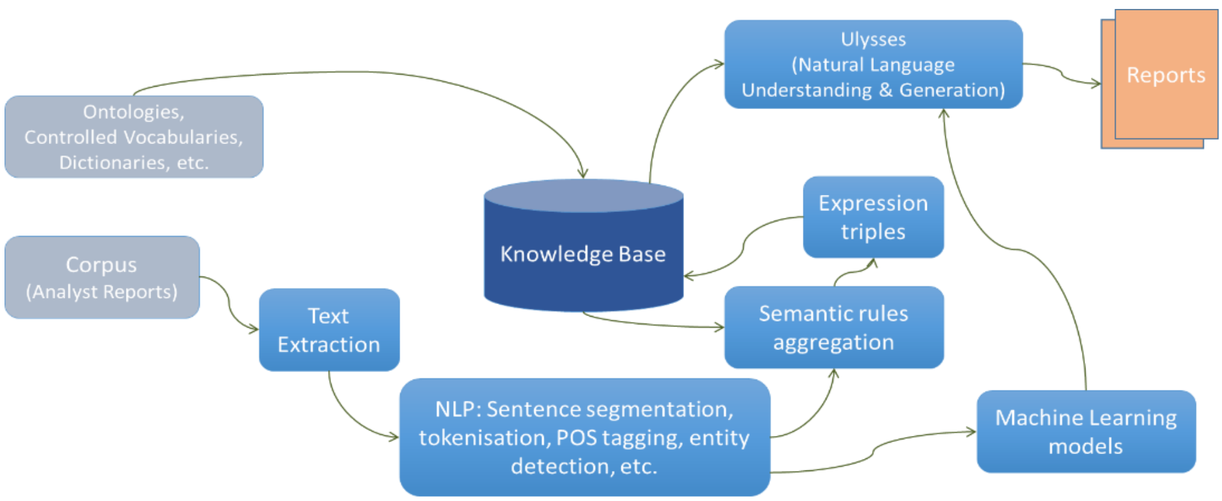
The engine has the capability to recommend a MAS grading for each customer consistent with its financial performance and MAS grading guidelines using machine learning techniques. Each borrower is assigned one of the five categories (either Pass, Special Mention, Substandard, Doubtful, or Loss) of MAS Grading based on their financial statements. [4]

The engine is also capable of indicating a potential credit deterioration flag as an early warning indicator for credit monitoring purpose, taking into consideration movements in leading macroeconomic indicators and the borrower's financial performance to date.

B) Credit Opinion Module:

In order for the rationale behind the proposed recommendation based on the financial matrices and trend performance of each customer can be articulated, a credit opinion module is built. It consists of different activities which supports the opinion report generation. (Figure 3)

Figure 3: Conceptual Framework of the Credit Opinion Generation Module

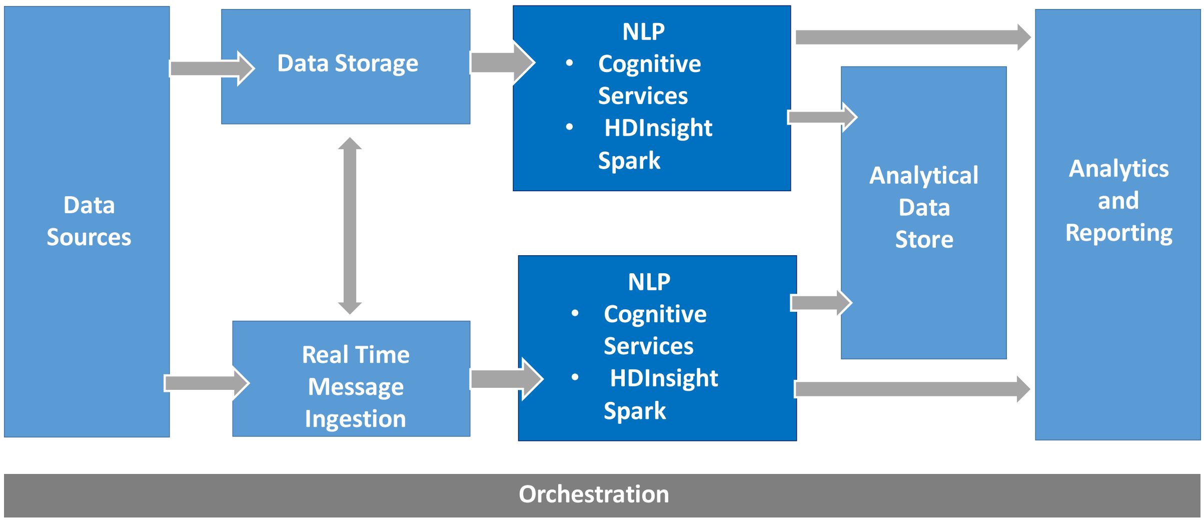


[4] When building the engine, its performance was evaluated using 8 years of historical data. It was trained using data for the first 5 years before subjecting the engine to a sample testing to evaluate its predictive capability next 3 years (2017-2020). The performance of its output has been positive, showing consistently high accuracy when it was backtested as well as subkjecting it to two rounds of blind tests.

The engine also consumes a series of market data such as GDP of leading advanced economies, US treasuries, interest rates and currency changes etc to compute an early warning risk indicator that will flag out a warning for the borrower should the engine "foresees" potential credit risk in the near-term horizon.

Data pipelines allows users to transform data from one representation to another through a series of steps. Data pipelines are a key part of data engineering and other data preparation activities. (Figure 4)

Figure 4: An Overview of the PoC Data Pipeline Structure



The system contains the automated process to extract the relevant data from source systems for historical grading, financial statements, and previous credit analyst reports etc that are all stored in the designated data storage. The real time message ingestion stores other data points, extracted from external data sources.

C) Natural Language Module:

There are three steps to process the relevant unstructured data used to train the engine that will generate the credit opinion for each borrower.

Step 1- Natural language processing (NLP): The development of the automated applications and services that can transform human languages to machine understandable language and act as per the user's requirements.

Step 2- Natural Language Understanding (NLU) The interpretation of the meaning the user intents to communicate. This step is responsible for understanding the intent of the data and identifies the relationships between different contexts given in the data.

Step 3- Natural Language Generation (NLG): The process of producing meaningful phrases and sentences in the form of natural language.

2.4. The Core Engine Design

In addition to the engines and modules created, to build the core engine, machine learning (ML) and deep learning (DL) libraries were incorporated to enhance the efficiency of text generation and achieve better results.

A) ML Techniques- Markov Chain Model[5]:

A Visual of a Markov Chain Generator

The first model is based on the "Markov chain generator" to build Markov models of large corpora of text, and generating random sentences from that. Markov chains are mathematical systems that hop from one "state" (a situation or set of values) to another. The Markov chain model consider respective corpus to train the engine, after which the engine can generate opinions randomly.

B) DL Techniques: Long short-term memory LSTM

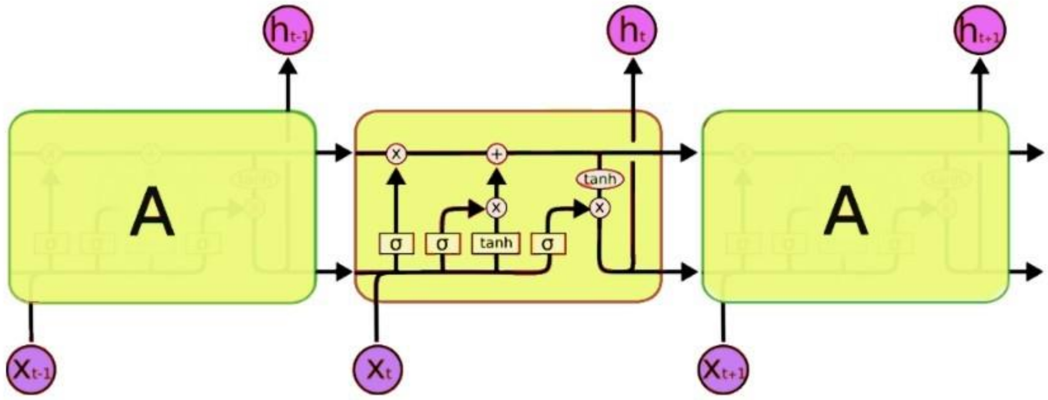
Deep learning techniques teach computers to "learn by example" like humans would naturally do. The core model the PoC uses is dense LSTM (Long short-term memory), a cutting-edge deep learning algorithm known to be one of the most effective solution for sequence prediction. Figure 5 gives a simple architecture of LSTM network[6]:

---

[5] https://chatbotslife.com/notes-on-remixing-noon-generative-text-and-markov-chains-84ff4ec23937

[6] http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Figure 5: An architecture of the LSTM Network



LSTM network is comprised of different memory blocks called cells. There are two states that are being transferred to the next cell; the cell state and the hidden state. The memory blocks are responsible for remembering things and manipulations to this memory is done through three major mechanisms, called gates. These gates are Forget Gate, Input Gate, and Output Gate.

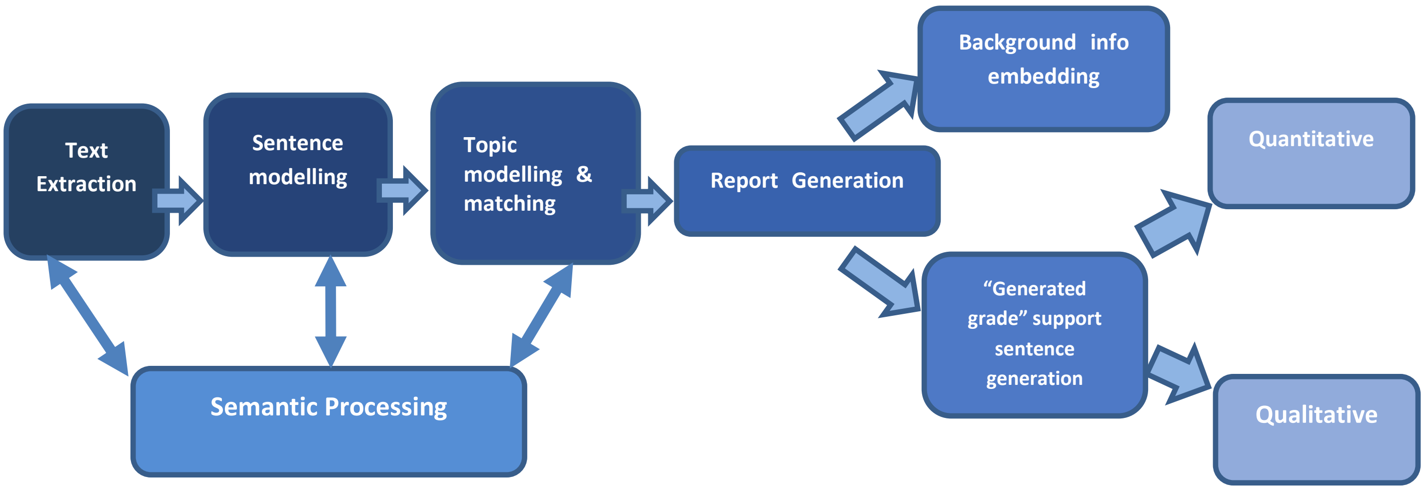
The Dense LSTM model is based on a sequential modelling paradigm with two LSTM layers, each having 400 units for processing. The first layer needs to be fed in with the input shape. In order to utilize the LSTM layers, added 400 units to process the same sequences of the vectorized data. The return_sequences parameter has been kept as “True” to learn from both the directions.

The fundamental idea is to start simple: Pass a text string as input along with a number of tokens the model is required to generate after the input text string. For example, if the user passes “How are” as the input text and specifies that the model should generate 2 tokens, the model will then experiment with various combinations and might generate “How are you doing” or “How are you man”, and so forth.

Combining all the above engines and modules is a core engine that consist of a data source unit (input unit) from where it collects the data to feed into the Transformation Unit, which has different components to perform a sequence of data engineering and training activities.

In particular, the **NLP Component** extracts the relevant data from the data storage and real time message ingestion unit Information stored include borrower's historical financial data, records of credit reports, proposed Grading assigned by the MAS Grading Recommendation Engine based on borrower's latest financial statement, records of credit reports and other company-related data from external sources.

Figure 6: Workflow of Natural Language Generation



The NLG component performs the actual analytics and reporting activity to generate the credit opinion after "understanding" all the plethora of structured and unstructured that is fed into the engine using dense LSTM (long short-term memory) model created.

3. Results of PoC in Generating A Credit Opinion:

The outcomes from the Grading Recommendation engine cumulate in the auto-generation of credit opinion in a form of a report using a deep learning NLG module developed for this PoC.

Besides a brief summary of the company's background and industry outlook and business performance using the data information stored in the pipeline, the engine is able to expose the rationale behind the MAS Grading recommendation. When evaluating the quality of the report generation, the following NLG metrics were used to assess the output:

- **Focus:** The generated opinion needs to have a focus and all the sentences need to contain information related to this focal point.

- **Structure and Coherence**: The summary should be well-organized and coherent body of information, not just a dump of related information. Specifically, the sentences should be connected to one another, maintaining good information flow.

- **Informative:** The opinion report should convey the key points of the context. The recommendation should be based on key points highlighted in the Grading engine and the system must elaborate the reasons to support the recommendations.

- **Non-redundancy:** The supporting text should not repeat any points, and ideally have maximal information coverage within the limited text length.

- **Referential clarity:** Any intra-sentence or cross-sentence references should be unambiguous and within the scope of the summary.

Based on all above NLG matrixes, the deep learning NLG module created for the PoC is able to produce an output that was fairly stable and consistent after just after 2 iterations.

There were challenges to overcome during the PoC. Both NLP/NLU/NLG techniques and semantic network models are nascent technologies that requires both time and data to achieve desired results. Combining both present even more challenges, which through time and constant iterations will get us to the goal.

The engine can be improved by providing a rich set of corpuses of expanded data points for rigorous training and learning. The language model can be upgraded to the latest technologies like BERT, Roberta etc. for additional accuracy and correctness with the appropriate infrastructural platform. Given time and data training, it will improve with each iteration to evolve from its current nascent stage to maturity.

4. Moving Beyond The POC

The PoC outcome suggests that the introduction of an engine equipped with basic cognitive ability similar to a human being would be a milestone towards in the development of explainable AI.

Not only will the engine produce a predictive outcome for human consumption, it will also be able to explain how it managed to derive at the conclusion. In our case, a credit opinion consistent with its recommendation logic backed by reasons.

The above-mentioned favourable outcomes from this PoC gives us confidence to bring this PoC to production. Envisaged high-level next steps for implementing this in production are as follows:

- expand current knowledge graphs both horizontally and vertically – ontologies and controlled vocabularies pertaining to industries, concepts e.g. macro-economic indicators, financial ratios, etc.

- introduce new algorithms e.g. machine learning, deep learning and reinforcement learning models and train them against wider and deeper data sets and refine models that has been established during the PoC phase.

- develop technical infrastructure i.e. both functional modules e.g. workflow, messaging, approval, etc. and non-functional capabilities e.g. scalability, security, performance, etc.

- deploy above mentioned components in an incremental phased approach with clear success metrics from each phase.

Moving forward, with the advent in financial technologies, it can be envisaged that credit risk evaluation would evolve into a process that is more time sensitive and predictive manner.

Copyright and Disclaimer