FSTI PROOF OF CONCEPT WHITEPAPER – SMART INTAKE WITH TRANSFER LEARNING

Zurich Insurance Company Ltd (Singapore Branch)

August 2021

Table of contents

1. Executive summary

Commercial Insurance's processes are manual, complex, case specific in nature, and have comparatively lower data volumes vis-a-vis the fast flow and more standardised processes for Personal Insurance. These factors have historically made Commercial Insurance less suitable for applying Artificial Intelligence (AI) / Machine Learning (ML) technologies for productivity / efficiency improvements.

The development in deep learning models that are based on Transformer architectures and leverages transfer learning, however, have demonstrated great potential to breaking down these barriers, and achieving promising level of performance even with low level of use case specific training data.

This Proof of Concept (POC) is commissioned with view to assess the feasibility to applying AI/ML models with transfer learning to facilitate the undertaking of a traditionally manual and complex Commercial Insurance process at the Singapore Branch of Zurich Insurance Company Limited (Zurich Singapore).

This whitepaper is comprised of the following sections:

- Section 1: Executive summary, to outline the structure of this paper;

- Section 2: Background, to set out:
    - Historic challenges to AI/ML adoption by Commercial Insurers;
    - Overview of the different ML approaches available to solving case classification and entity extraction tasks, and why transfer learning is the most suitable solution for addressing the problem at hand;
    - Basis for selecting the specific transfer learning algorithms used at the pilot, and the adopted approach for pre-training / training;

- Section 3: Use case, to provide context to:
    - As-is process under consideration;
    - High level design of the to-be process that leverages the output of the AI/ML models with transfer learning;

- Section 4: Performance, to present:
    - Approach adopted to score and assess the technical performance of the algorithms in use;
    - Performance results observed;

- Section 5: Challenges, to sharing the key difficulties observed and lessons learnt in the areas of:
    - Handling of multi-label cases;
    - Handling of class imbalances;
    - Legacy process and system integration;
    - Remote collaboration during COVID-19;

- Section 6: Conclusion, to summarise:
    - Potential benefits, and risks to note from a broader implementation of the solution;
    - Strength and weaknesses of the solution, and the envisaged research and development areas;
    - Next steps for Zurich Singapore.

Lastly, overview of the project's governance and approach can be found in the Appendix for reference by other AI/ML practitioners to conduct similar experiment.

2. Background

2.1 Historic challenges to AI/ML adoption by Commercial Insurers

AI has great potential for the insurance industry. Insurance carriers are processing tens of thousands of data points a day to facilitate business critical decisions to be made in claims and underwriting. AI/ML technologies have proven that in many instances they can detect patterns in large pools of data that are very difficult to spot for humans. In many areas, AI/ML technologies have already surpassed human level performance today. For example, the ability for automated object detection and image understanding has greatly improved over the past ~10 years due to AI. At the ImageNet competition, an annual tournament for image understanding and machine vision, AI algorithms has outperformed humans since 2015[1]:

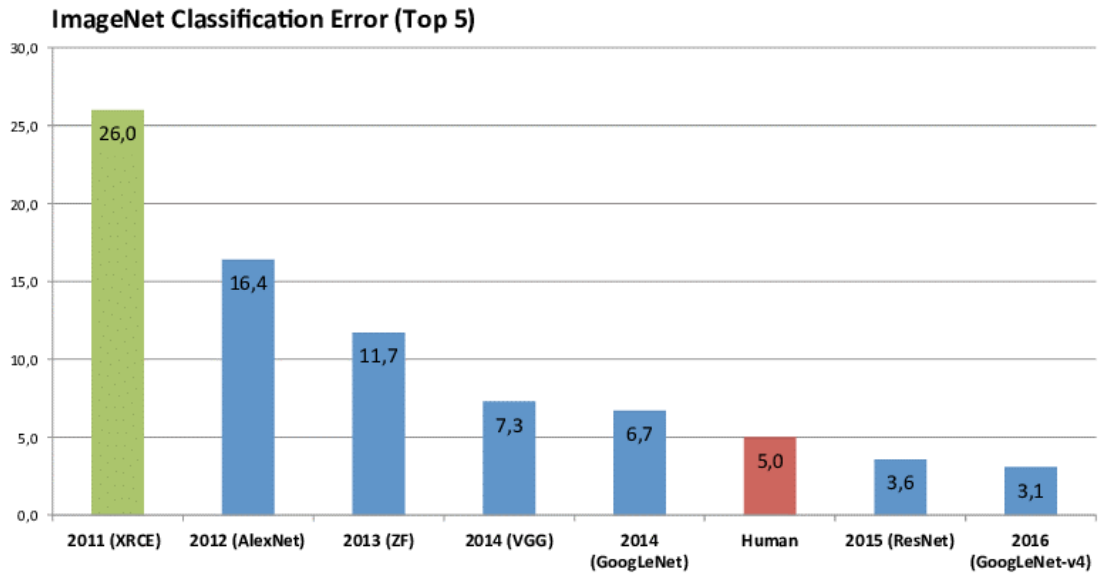

Image 1: Progress of AI (vs human) at ImageNet competition over time

However, the adoption of AI/ML in the insurance industry has so far focused mainly on Retail Insurance use cases since the segment has more data available, which is critical to training algorithms, and the processes are less complex compared with those in the Commercial Insurance space.

[1] *G. von Zitzewitz (2017) Survey of neural networks in autonomous driving*

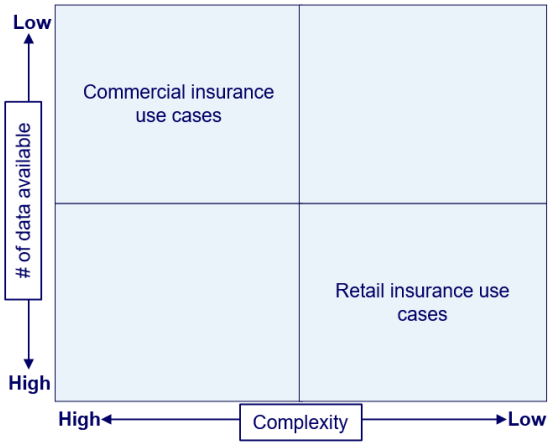

Image 2: Comparison of complexity and data availability between retail insurance and commercial insurance

Therefore, the adoption and use of AI/ML in Commercial Insurance has not taken off as of now. Echoing the comments raised at recent report published by Temasek[2], the road to AI is not without challenges and risks, and insurance companies cannot afford to use this as an excuse to shy away. Breakthrough in technology to lower the barriers to AI adoption by Commercial Insurer, especially in reducing the number of sample data required for effective ML, is truly welcoming news – There is significant opportunity for the incumbents to assess suitable use cases, and commence the journey to harness AI's full potential.

2.2 Machine learning approaches

There are different kinds of ML algorithms available to solving different problems. The focus of this whitepaper is placed on outlining the most suitable approaches to facilitate the execution of text classification tasks (i.e. assigning cases to the correct category) and entity extraction tasks (i.e. identifying and pulling specific parameter such as a name or address from a document / text block) as outlined in the diagram below.

[2]*https://www.temasek.com.sg/en/news-and-views/stories/future/ai-ethics-governance-heralding-new-era-financial-services*

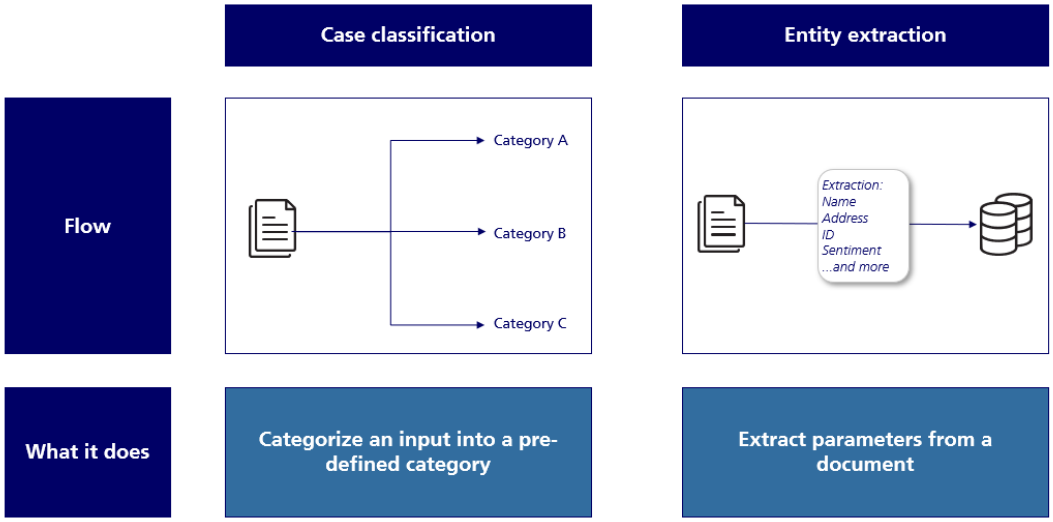

Image 3: Illustrative workings of Case Classification and Entity Extraction in the context of AI/ML

Other approaches in ML such as unsupervised ML are not covered in the report because they are not used for automated text processing, but are more suited for solving other data analysis work such as segmentations and clustering.

To process and understand text data, three types of algorithms / approaches are commonly being adopted:

1) *Memory based approaches*: These algorithms do not perform any advanced inference on the data but conduct a simple word count to assign a certain category based on frequency. For example, an inbound customer email is assigned into the categories "claims" vs. "underwriting" based on a comparison of the word count for both categories by counting and comparing claims related words such as "damage", "claim" with underwriting related words "new business", "risk rating", etc. Here, the Term Frequency - Inverse Document Frequency approach (TD-IDF) is one of the most widely adopted algorithms used for such tasks.

2) *Traditional machine learning approaches, a supervised ML models*: For the purpose of this whitepaper, we define "*traditional machine learning approaches*" as models and methods that do not use a large-scale transfer learning scheme, nor adopt the modern transformer architecture. The key difference between the algorithms mentioned here vis-à-vis the approach mentioned in Approach 1) is that the algorithmic inference goes well beyond a simple frequency / ratio-based approach.

   Traditional ML models are trained and calibrated using labelled training data with both the input (e.g. the email that is received by a broker) and the associated expected output / label (e.g. specific customer request such as "cancellation" vs. "renewal") pre-defined. The algorithm is expected to work by identifying the patterns that link the input and output, essentially to look for correlations and associations between the input and output. There are typically three ML algorithms used to conduct this task:

a. *Linear classifiers*: Algorithms that can be trained on data to identify linear relationships in the data, e.g. support vector machines.
b. *Tree based classifiers*: Algorithms that use an ensemble of decision trees to conduct classification tasks such as random forests.
c. *(Traditional) Neural network-based approaches*: Algorithms that use a traditional neural network architecture to conduct classification tasks such as recurrent neural networks and LSTMS – long short-term memory approaches.

3) *Modern transfer learning / transformer-based approaches, a pre-trained ML model*: Being significantly different from the traditional based neural networks that are set out under Approach 2), these are the "next generation" neural network architectures that can apply transfer learning, at scale and at speed over longer text sequences, and thereby making it a valuable new tool in Natural Language Processing (NLP) for use by data scientists[3]. The Wikipedia definition of transfer learning is:

> *"Transfer learning is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem. For example, knowledge gained while learning to recognize cars could apply when trying to recognize trucks."*

Models that use transfer learning (i.e. pre-trained on larger but non target specific data sets) exhibit a range of positives attributes vis-à-vis traditional models, including the need for less training data, more robustness in prediction results and reduced lead time to training the algorithms. However, the exact training data set, type of pre-training, use case in question, etc. can influence the success and viability of this method.

Transfer learning can be compared to the way human learning works as humans leverage accumulated knowledge over various tasks rather than requiring to learn each new task from scratch. Humans build up knowledge and experience over time and this enable us to make the right decision without requiring sight of lots of examples to get to the right conclusion, e.g. we can apply the reading and writing skills leant at school to allow us to perform our jobs in current day, without requiring the need to learn on the job from scratch on how to read a document by looking at tens and thousands of examples. The analogy between human learning and ML are far from perfect, but the key point to note here is that transfer learning through knowledge accumulation enables a faster and easier grasp of

[3] *https://deepai.org/machine-learning-glossary-and-terms/transfer-learning and https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a and http://nlp.seas.harvard.edu/2018/04/03/attention.html and https://www.analyticsvidhya.com/blog/2019/06/understanding-transformers-nlp-state-of-the-art-models/*

related tasks and challenges for problem solving.

In the remainder of this section, we will outline why Approaches 1) and 2) may not be adequate and fit for purpose to deliver meaningful automation in the Commercial Insurance space and how the models outlined in Approach 3) can provide a true step change to the AI/ML enabled automation in Commercial Insurance.

In ML, the volume of training data and the performance of a solution usually correlate positively with each other until reaching a point of diminishing return that the provision of additional data point yield minimum marginal benefit.

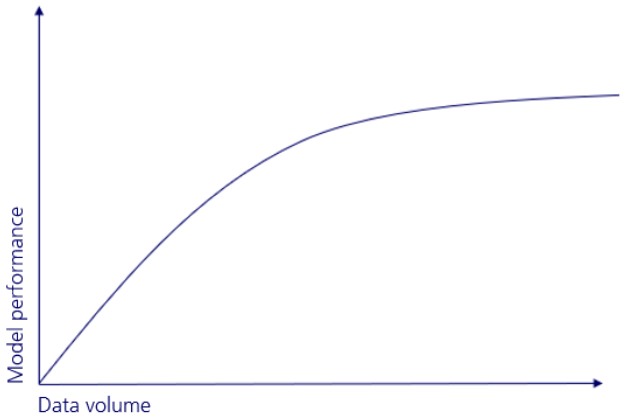

Image 4: Illustration of the relationship between the provision of additional labelled data from a target process / task and the associated ability and quality of the performance of an algorithm to perform the task

One of the biggest draw backs of the count / scoring based approach is that the decision of the associating models is derived based on single parameters such as simple sums, ratios, or distance measures, etc., for solving simple problems with small data sets[4], thus these algorithms are generally not used for complex tasks and inference.

For traditional ML algorithms, though they do not have the above-mentioned constraints in terms of data volume and can support more complex inference[5], they suffer from a different constraint, i.e. the need for large amount of data to learn and identify patterns from the data. For example, the popular Support Vector Machines are required to be trained using usually thousands of labelled data before it can adequately perform binary classification tasks on text data; The more complex neural networks are often trained on tens of thousands or even hundreds of thousands of labelled data before they can

[4]*https://towardsdatascience.com/k-nearest-neighbors-knn-algorithm-23832490e3f4#:~:text=5.-,Limitation%20of%20KNN,memory%20used%20by%20the%20algorithm*

[5]*This broad statement can vary based on the type of algorithm used but does describe the vast majority of these algorithms.*

perform their assigned task (well)[6]; For very complex use cases such as self-driving vehicles, millions or even billions of training data are needed[7].

This means that significant work is required to gather and label data, which in turn would increase the effort and cost of the ML projects. In addition, it is not always the case that sufficient labelled training data is available in the first place for use. This prohibited the adoption of ML in many "*data poor*" functions or industries. Although there are approaches to allow for the use of pre-trained components, none of these models can achieve the speed nor support the scale to truly deliver a big step change until the advent of large scale, transformer based deep learning models over the past 3 – 4 years with Google's BERT being one of the most well-known models used for NLP tasks in this category.

Bidirectional Encoder Representations from Transformers (BERT) is a deep learning architecture that was published in 2018. There are different sizes of BERT, but the standard version of BERT (M) was pre-trained on a book corpus (a dataset consisting of 11,038 unpublished books from 16 different genres), and the English Wikipedia (with text passages containing 2,500M words)[8]. This pre-training of data, alongside a modern learning and processing architecture has yielded impressive results, including:

- A new record setting result on the General Language Understanding Evaluation (GLUE), i.e. a benchmark that compares the text understanding competencies of algorithms to a human baseline by measuring the reading comprehension on a set of tasks, with BERT being the first algorithm ever to achieve a reading comprehension score that was close to the human level in 2018[9] albeit later releases of similar model architectures that were inspired by BERT outperformed the human level performance on this standardised test[10];
- An improvement of search results across various languages as Google decided in 2019 to include BERT to power the search in over 22 languages, given the

---

[6] *https://hackernoon.com/why-does-machine-learning-require-so-much-training-data-cc839cd62fa5 and https://journalofbigdata.springeropen.com/articles/10.1186/s40537-014-0007-7 and https://ch.mathworks.com/de/discovery/deep-learning.html*

[7] *https://blog.cloudfactory.com/autonomous-vehicle-training-conundrum and https://towardsdatascience.com/teslas-deep-learning-at-scale-7eed85b235d3*

[8] *https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/ and https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/ and https://github.com/google-research/bert and https://towardsml.com/2019/09/17/bert-explained-a-complete-guide-with-theory-and-tutorial/*

[9] *https://medium.com/syncedreview/best-nlp-model-ever-google-bert-sets-new-standards-in-11-language-tasks-4a2a189bc155 and https://www.wired.com/story/computers-are-learning-to-read-but-theyre-still-not-so-smart/ and https://gluebenchmark.com/*

[10] *https://medium.com/syncedreview/best-nlp-model-ever-google-bert-sets-new-standards-in-11-language-tasks-4a2a189bc155 and https://www.wired.com/story/computers-are-learning-to-read-but-theyre-still-not-so-smart/ and https://gluebenchmark.com/*

strong performance of BERT[11]; and
- A better understanding and analysis of complex patent data[12].

In addition to BERT, other market players such as Facebook have released similar model architectures in recent years, e.g. RoBERTa, and Google has also released variants of BERT such as ALBERT to provide a more "lightweight" / faster architecture for BERT[13]. These developments have made the application of AI/ML enabled automation to "data poor" processes a possibility – A game changer.

2.3 Choice of the transfer learning algorithms selected, and pre-training / additional training applied

For this POC, Google BERT[14] was used for the case classification work while Google ALBERT[15], which is a distilled version of the original BERT model, was adopted for the entity extraction work. Google BERT was chosen because it is the oldest of these "*new breed*" models and, thus, there was a lot more documentation and information available vis-à-vis other transformer-based models, which were just released a few months ago. ALBERT was chosen, because of its good performance on question / answering (i.e. entity extraction tasks) on external benchmarks such as SQUAD – Stanford Question Answering Data Set.[16] The outcome that we have achieved with these two models will be detailed in Section 4.

To prepare the algorithm for the POC, pre-training was carried using approximately 90,000 documents from the following three sources:

- *General email files*, to enable the model to better recognise text structures and blocks in emails;
- *General Commercial Insurance documents*, such as underwriting policies, training manuals, and other documents that contain Commercial Insurance related subject matter specific language, to guide the model to better pick up certain word types, word order, word sequences and other latent features; and
- *Unlabeled data samples (e.g. underwriting submissions in the format of email correspondence, document attachments, etc.) from the existing process of the*

---

[11] *https://blog.google/products/search/search-language-understanding-bert/*

[12] *https://cloud.google.com/blog/products/ai-machine-learning/how-ai-improves-patent-analysis*

[13] *https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/ and https://ai.googleblog.com/2019/12/albert-lite-bert-for-self-supervised.html*

[14] *https://github.com/google-research/bert*

[15] *https://github.com/google-research/albert*

[16] *https://venturebeat.com/2019/09/26/google-ais-albert-claims-top-spot-in-multiple-nlp-performance-benchmarks/#:~:text=On%20the%20Stanford%20Question%20Answering,gets%20a%20score%20of%2089.4%25*

use case that are intended to be carried over to the target process, to assist the model to become more familiar with the language, format and structure of the input it is expected to process in the future.

These transfer learning efforts on both documents and text strings are key factors that can influence model accuracy success.

Beyond these pre-training, the following additional tuning and modeling steps are applied:

- *Additional training and tuning of the model*: This involved the manual collection of applicable cases, pre-processing of these cases by way of parsing of the gathered emails and attachments, and preparation of labelled data. Here, tuning is achieved by way of identifying the optimal combination of hyperparameters to maximise the model's performance using grid search.

- *Applying additional business logic and configuration to the model*: Here, business rules and supplementary data (e.g. Singapore's bank holiday dates) were introduced to the model to facilitate the orchestration of the process, such that it can recognise and manage the relevant process steps (in this case the assignment of due dates on business days) that cannot necessarily be trained using historic data.

3. SmartIntake – Pilot study to assess the feasibility to apply AI/ML models with transfer learning to Commercial Insurance processes

3.1 Objective

The emergence of the transfer learning models, and associating ML architectures meant that the AI can now be trained using a small(er) amount of business case specific data – This has opened up the opportunity for automation to be introduced to support Commercial Insurance's "*data poor*" processes.

With this in mind, the Zurich Singapore commissioned a pilot study, code name SmartIntake, to assess the feasibility to apply AI/ML in the use case to (semi-)automate the creation / update of CRM system entries by using information contained in inbound emails (with attachments) from internal (e.g. underwriters) and external (e.g. brokers) stakeholders, with aim to improve the efficiency of a highly manual and complex Commercial Insurance process.

3.2 As-is process under consideration

In Commercial Insurance, the interactions between the insurance carrier and the insured / customer are mainly carried out via the Brokers, and the instructions and associating information are shared mostly via emails.

High level overview of the existing business process that is supported by the Underwriting Service team (UWS) is as follows:

1) Underwriter receives an email, containing single or multiple service requests for new or existing customers, from Brokers or other partners;
2) Underwriter evaluates whether s/he can / should handle the request directly, or can be performed by the UWS;
3) If the task is part of the UWS' defined scope of responsibilities (e.g. to support the execution of the operational / administrative aspects of the underwriting / customer servicing processes, e.g. carrying out sanction check, update of contact details, etc.), the Underwriter will then forward the received email to the UWS shared mailbox, along with his / her instruction on the handling of the requested task, and supplement the email with additional information, if required;
4) All requests received at the UWS shared mailbox are triaged, with workflow tasks subsequently created and assigned to specific onshore / offshore teams or individuals, and logged at the CRM application along with an expected due date for activity status tracking;
5) After the task is performed, the status of the workflow task is updated at the CRM application, and that the Underwriter, and if required, the Brokers / customers, is then informed of the completion status of the task.

The above outlined process is highly manual, and three pain points are typically observed:

- *Bottleneck and repetitive nature of the work*: As the end-to-end process is

entirely manual regardless of the complexity of the requested task, and all of the emails / attachments need to be read by the assigned task handler one-by-one, the lead time to complete the case / update the associating status can be long, and bottlenecks can be created especially during peak business seasons. Some elements of the process are also highly repetitive, and thus not necessarily an enjoyable / mentally challenging undertaking for employees.

- *Long training lead time and repeating need for staff on-boarding*: One of the unfortunate byproducts of the above pain point is that the risk of staff attrition is high. As the process is highly manual and varied, and the handling of certain tasks can be complex, the lead time and thus effort to onboard new joiners is high.

- *Limited end to end analytics*: Whilst the existing process allows for high level operating metrics, e.g. number of cases handled, to be captured, there is limitation to provide granular data insights, e.g. average duration to handle a property case when all information is provided by the broker in one email.

3.3 High level design of the to-be process

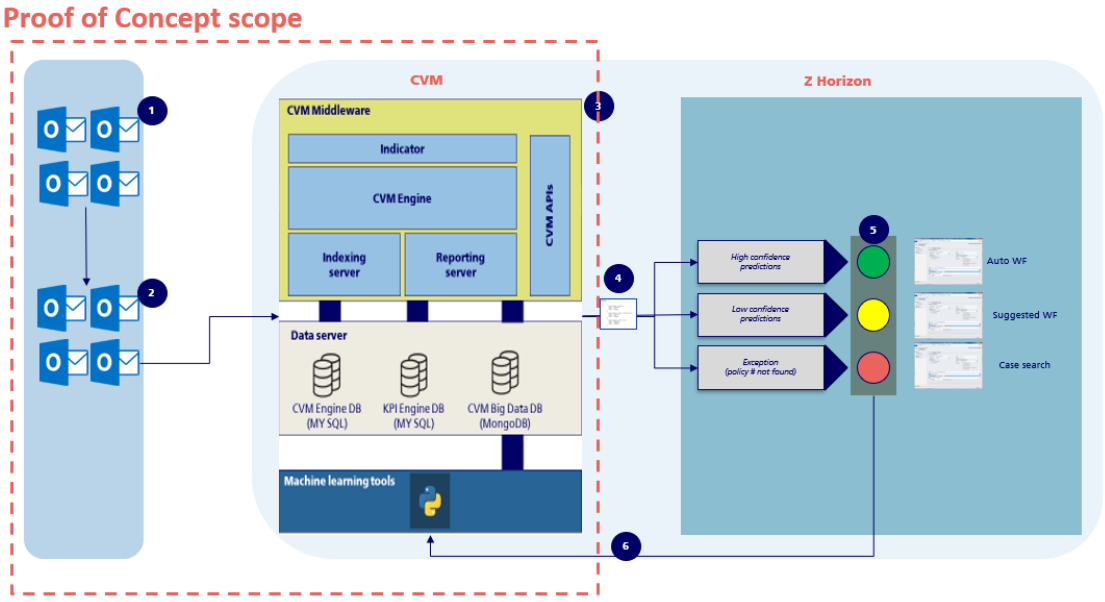

Image 5: Illustrative overview of the reengineered end to end process, and focus of this whitepaper

There are 6 steps involved in the proposed to-be business process[17] for the use case to automate the creation / update of CRM system entries by using instructions and

[17]*The write up of the to-be process has been simplified to exclude the exhaustive details i.e. required to be set out in the solution design to allow for the handling of process exceptions, validation checks, etc. at the CRM systems in the processing of the case based on the provisioned instruction and content.*

information contained in inbound emails – At high level:

1) The Underwriter forwards the received email to the UWS shared mailbox, along with his / her instruction on the handling of the requested task, and supplement the email with additional information, if required.

2) The email is automatically routed to a shadow email inbox from where it is extracted by an Application Programming Interface (API) and placed into the cloud instance as a case, where the AI engine resides.

3) The case goes through a processing pipeline, with the AI engine performing the following:

    a. The email and any accompanying attachments are read;
    b. The classification of the case is determined (i.e. what the email is about); and,
    c. Key parameters from the email and attachments (e.g. due date, customer name etc.) are extracted,

    after which these case classification and entity extraction outputs are stored in a database along with their corresponding confidence score[18] from the ML algorithm.

4) In addition to the storing of the output, subject to the nature of the case classification identified, appropriate choice of the downstream processing logic is determined before the relevant API is called to sending the relevant extracted entity parameters to the CRM system in a structured format for processing.

5) At the CRM system, users can then review the AI engine processed case's traffic light indicator to assess whether manual intervention is required, with:

    a. Green traffic light, indicating that:
        i. all information that was required for the processing of the identified case was found;
        ii. the case is processed with high confidence by the AI engine; and
        iii. human intervention is not required.
    b. Amber traffic light, indicating that:
        i. all information that was required for the processing of the identified case was found;
        ii. the case is processed with low confidence by the AI engine; and,
        iii. manual review of the machine derived output is recommended.
    c. Red traffic light, indicating that an exception case has been identified requiring manual review and intervention to complete the processing of the case at the CRM system. There are several types of cases that are

[18] *Confidence scores are an automated probability value produced by the AI engine that provides information about the level of certainty of the prediction task.*

expected to be flagged by the AI machine as ”Red” cases - Two of the most common examples are:

i. Attachments cannot be processed, i.e. when an attachment is not machine readable or password protected; or
ii. Missing information, i.e. when critical information (e.g. customer information) is missing from the case.

6) The results from the manual intervention are periodically sent back from the CRM system to the AI engine. This feedback process enables a periodic re-training of the AI algorithm, which is particularly relevant for “Amber” cases.

The workings of the core of these steps, especially the AI engine’s interaction with the CRM system, are specific and proprietary to Zurich Singapore. However, we will look to share our findings on the technical performance result of the AI engine and our learnings in the upcoming section of this whitepaper, which we hope will be of interest and benefit to fellow data scientists who are looking to solve similar problems.

4. Technical performance results

4.1 Approach for assessing the technical performance of algorithms

Performance of the algorithms for executing the case classification tasks are assessed using confusion matrices. Take the processing of a binary prediction task, as an example, confusion matrices work by determining how many cases an algorithm has correctly classified by comparing predicted and actual values, and grouping them into the following categories for analysis and review[19]:

1) *True positives*, indicating that the result was predicted by the model, and it was correct;
2) *True negatives*, indicating that the result was not predicted by the model, but it was correct;
3) *False positive*, indicating that the result was predicted by the model, but it was incorrect (Type 1 error);
4) *False negative*, indicating that the result was not predicted by the model, and it was incorrect (Type 2 error).

These four categories are shown below in a two-by-two (2x2) confusion matrix with the "predicted label" representing the machine's prediction and the "actual label" representing the ground truth derived from the manually prepared label.

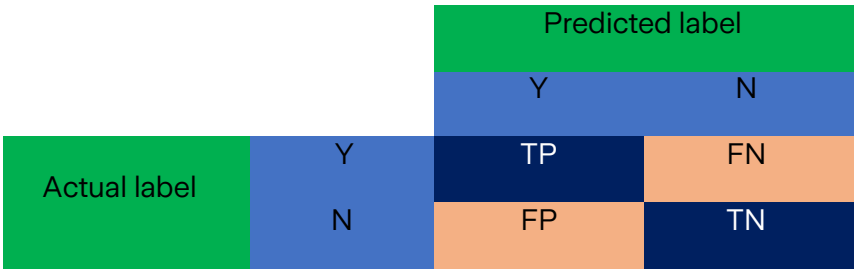

Image 6: Illustration of the workings of a 2x2 confusion matrix for a binary prediction task

For entity extraction, however, use of confusion matrices for measuring the performance of the algorithms are not fit for purpose because every single word could be considered as a dimension. Thus, a case with 500 words would create a 500x500 confusion matrix, which is too big to outline any meaningful insights. Therefore, a simple accuracy metric is used instead for measuring the number of correct predictions divided by all predictions for a certain defined parameter. For example, if we extract the underwriter's name 10 times and it is seven times correct, then the accuracy would be 70%.

[19] *https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 and https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html*

4.2 Performance results

Whilst the 2x2 matrix is a good way to demonstrate the workings of the confusion matrix, the performance of the case classification task by the algorithm on the use case, involving both single label and multi-label scenarios[20], is assessed across many more categories.

The below diagrams illustrate the confusion matrices of the case classification results for single label prediction runs:

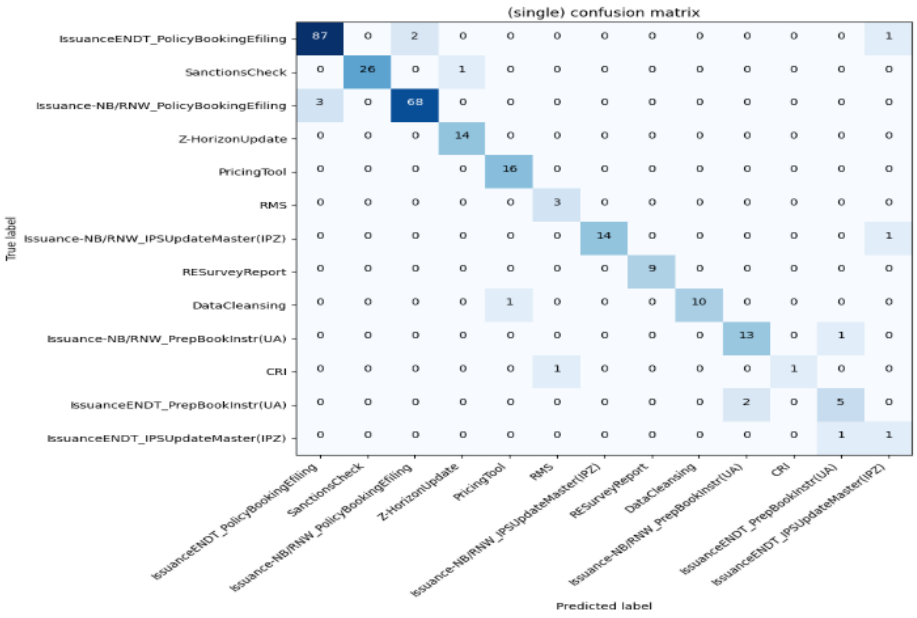

Image 7a: Result after the first training run with 3627 training dataset - 95.7% accuracy (correct cases divided by all cases)

[20] *Single label = one email has just one associated workflow task; Multi-label = one email has multiple workflow tasks.*

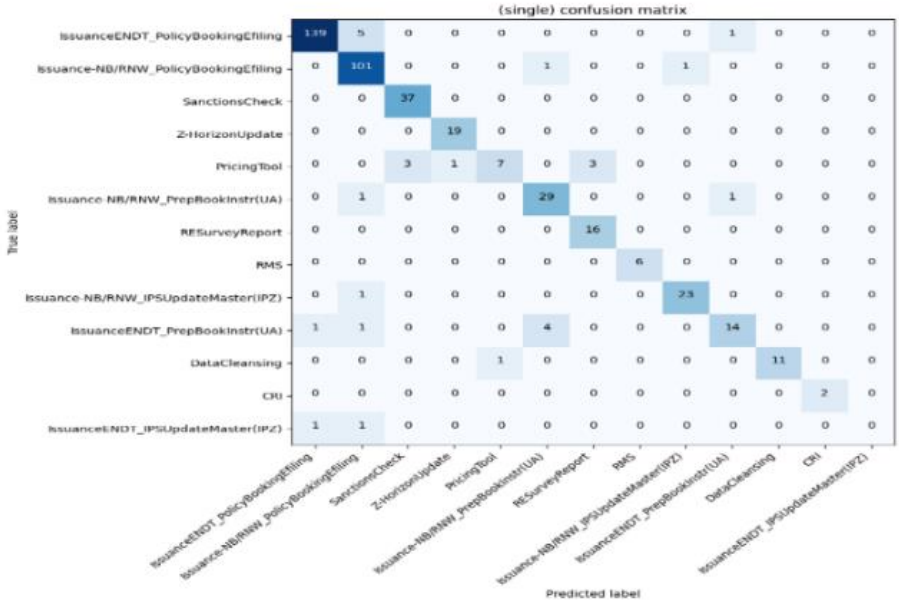

Image 7b: Result after final training run with 5581 more challenging training dataset - 93.6% accuracy

All in all, for single label cases, the performance observed is highly encouraging and positive, with a prediction accuracy result (i.e. the number of correctly predicted cases divided by all cases) of over 90%. However, it also showed that the results are better for high frequency categories vis-à-vis categories with smaller samples. Results for the second run (see Image 7b) were slightly lower vis-a-vis the first run (see Image 7a) despite more training data. The reason for this is threefold:

- Firstly, for the second run, we collected relatively more data for the low frequency categories to improve the performance on such low volume cases, thus making the data set more challenging for the algorithm as data frequency and prediction performance positively correlate.

- Secondly, we added one additional category, growing the categories from 13 to 14, which made the overall prediction task for the algorithm even more challenging.

- Finally, small deviations in performance scores are also driven by random chance.

The below diagram illustrates the confusion matrix of the case classification results with multi-labels:

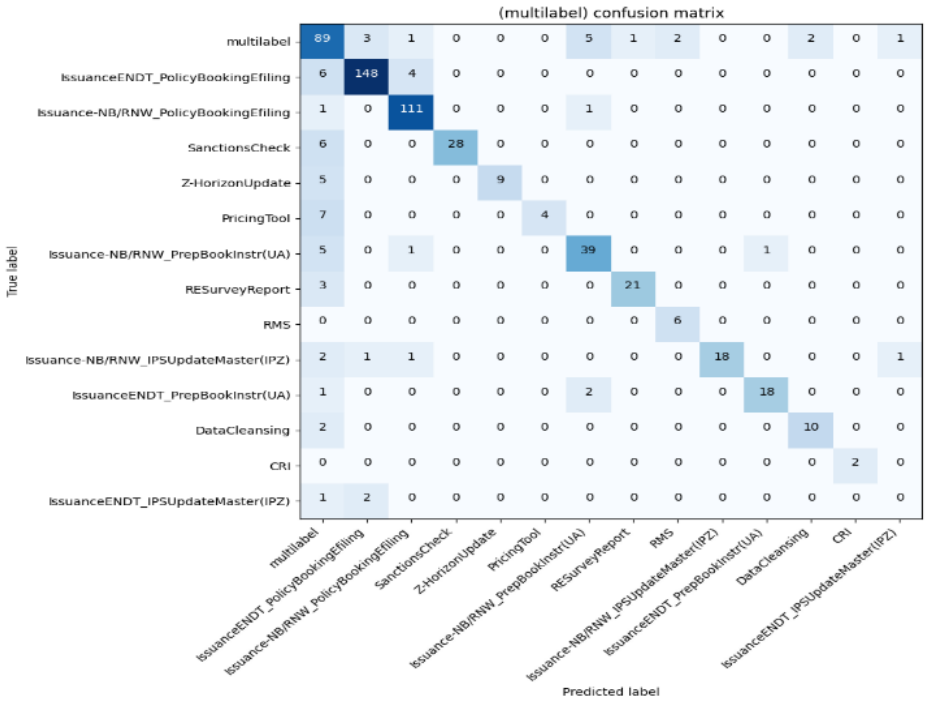

Image 7c: Result after the final training run was 81.67%.

Here, the prediction accuracy result weas lower due to the multi-label use case' complexity. This will be further examined in Section 5.1.

	Correctness	# Test Cases
account name address	0.70588235	51
assignee	0.83098592	71
due date	0.8	10
policy id	1	30
uw year	0.96078431	51
technical lob id	1	19

Image 8: Illustrative performance result for sample of the entities extracted after the final test run

For the entity extraction, without weighing the result of all of the extracted entities' frequency of occurrence but taking the mean of the accuracy percentage values, a balanced accuracy score of 83% was achieved.

Here, it should be noted that it is not uncommon for the performance result of the model to differ slightly from the pre-go live training runs after a period production run because it is unlikely that they will share the same data distribution. For example, in the training dataset, a certain email category, e.g. endorsements, may account for 10% of all of the data samples, but the share of such category may be somewhat higher or lower under production settings. Thus, the overall prediction accuracy may change upwards or downwards depending on whether the specific category's average accuracy is above or

below the mean.

At the point of writeup of this whitepaper, the end-to-end solution remains under testing but the project team is confident that the good results observed from the training run could be replicated with the production data “in the wild”. As retraining takes place over time using the tracked production data and results, it is expected the model’s accuracy and performance will continue to improve.

5. Challenges and lessons learnt

Building and integrating ML technologies into complex business processes is a very challenging activity even when using the latest available tools and algorithms. Key challenges observed during the implementation of the project are outlined in this section.

5.1 Handling of multi-label cases

Multi-labels are defined as "*a variant of a classification problem where multiple labels may be assigned to each instance*"[21]. For example, an email from a broker could include not one, but multiple requests such as requesting a price quote and enquiring about the status of an earlier submission. As noted in empirical findings, ML algorithms often struggle to identify all of the relevant categories for case classification, as well as to correctly predict all of the outcomes for the multi-label cases.[22] Our work has confirmed these findings as the performance for multi-label cases was observably lower than the single label (one email with one category / task) cases.

Various approaches were tested by the team to overcome this challenge, such as adjusting certain components of the ML model. However, results remained well below the single label cases despite these model adjustments. Consequently, for the first release of this solution, these will be flagged for human review whenever there were hints in a case that it belongs to the multi-label case category.

5.2 Handling of class imbalances

Data sets are imbalanced when "*observations in one class are significantly higher than the observations in other classes*."[23], or more simply put, the classes are not represented equally. This is a problem because in imbalanced data sets, ML algorithms tend to focus mostly on the largest case categories in the learning process, and disregard the smaller categories[24]. Thus, whilst the performance for the larger categories may be good, the accuracy scores for the smaller categories may exhibit substantially poorer results. For example, the case classification training data for this POC were initially highly skewed

---

[21] *https://en.wikipedia.org/wiki/Multi-label_classification and https://machinelearningmastery.com/multi-label-classification-with-deep-learning/*

[22] *https://journals.sagepub.com/doi/full/10.1177/1550147720911892*

[23] *https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/*

[24] *https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ and https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/*

with the top 4 categories accounted for over 86% of the overall gathered samples - Here, whilst the two largest categories were exhibiting performance score of over 96% accuracy, the result for the two smallest categories with the lowest data volume was initially at 50%.

To improve the model's results in a balanced fashion, the sampling for those smaller categories was adjusted. More specifically, extra sets of the smallest categories were added to the training data to growing the share of these low frequency categories by way of retrospective oversampling. This has improved the individual results, for example, for CRI from 50% to 100%, for Data Cleansing from 90% to 100%, etc.

5.3 Legacy process and system integration

Beyond the above outlined ML challenges, it is important to note that the potential complexity in the design, build and test process to ensure that the ML outcome can be seamlessly integrated into the underlying business processes and applications should not be under-estimated.

This complexity will vary between companies as no companies' way of working are the same. Only when the end-to-end integration is achieved, by way of seamless handoffs between the API over the defined business orchestration logic, to matching the extracted data to the right policy and / or workflow, can the envisaged business benefits of the automation be truly realised.

This can be achieved with diligent assessment of the business requirement, review in the workings of the APIs across applications and detailed definition of the transaction flow, tireless testing and finetuning of the solution, etc. and especially close collaboration, support and communication between the business, application and project teams.

5.4 Remote collaboration

A final challenge to note in this project was that it was made up of team members from three countries and time zones, and across two continents, who had to work and collaborate together yet unable to meet in person due to the COVID-19 pandemic.

To overcome this challenge, remote collaboration tools such as Microsoft Teams, SharePoint and other tools were used to enable simple engagement and interaction between the teams. Whilst the project was able to be run this way with little loss in productivity and efficiency, certain activities such as end-to-end process reviews, and requirement workshops, etc. had to be extended or repeated to achieve the right level of mutual understanding and cross-team collaboration to finalise the relevant deliverables.

To fully overcome the challenges outlined in Section 5.1 and 5.2, further advances in ML research are needed, which will be explained in the next chapter.

6. Conclusion and outlook

This whitepaper has outlined how Zurich Singapore has worked to apply ML solution in a traditionally highly manual Commercial Insurance process, the performance of such solution, and also the challenges and lessons learnt it has experienced.

In this final part of the paper, we will further examine the following three questions:

1) What are the potential benefits, and risks of adopting this technology at a broader scale?
2) What has this project told us about the strength and weakness about today's ML solutions, and the envisaged key research areas to drive improvement?
3) How should Zurich Singapore continue to evolve and improve its ML solution and capability?

6.1 Potential benefits and risk from a broader adoption of the solution

The use of AI/ML algorithms to enable automation in Commercial Insurance creates both new opportunities and risk. The key benefits from a broader adoption include:

- *Higher productivity, and better customer and employee experience*: Automation improves not only the turnaround time and processing efficiency, but also allowing capacity to be freed up to provide a deeper degree of personalization support, higher quality of service, and in turn provide a fuller and richer experience to the customers – This is highly important for companies to stay competitive, and also highly relevant for employees to enrich their skillset and improve their work satisfaction from the processing of manual / repetitive tasks to leveraging AI/ML to deliver a higher quality output, and allow for more time to be spent in delivering higher value add services that makes a difference to the company's proposition.

- *Higher resilience and flexibility*: COVID-19 has demonstrated how external shocks can create significant risks to business models and business processes. If an organization is able adopt a digital business model whereby AI/ML algorithms can take on more of the work, this can in turn make the business less vulnerable / more resilient to natural events such as pandemics or extreme weather conditions.

- *Better analytics and continuous improvement*: A fully automated process allows for the generation of much richer analytics and insights as every processing step (e.g. case received, case reviewed, case closed etc.) will have a time stamp, and in turn can be tracked and monitored. Furthermore, with feedback loop built in to enable the AI/ML to be trained, continuous improvement to performance can also be assured.

Some examples of the emerging risks to note are as follows:

- *Cyber risk*: The more critical business activities rely on algorithms, the higher the risk that a cyber-attack, or some other IT problem, may create a significant

disruption to the business processes. Thus, it is key to establish robust IT security capabilities and disciplines, plan for redundancies and back-ups, and be ready to fall back to manual processes, etc. to safeguard business continuity and mitigate the risk / impact of extreme shocks.

- *Knowledge and skill attrition risk*: As algorithms take on more and more elements of the business process, there is a risk that critical understanding of the why / what / how certain steps and controls are required to be carried out may be lost, especially as:

    a. The older / more experienced generations who had an intimate understanding of the workings of every manual steps start to retire, and that the younger / less experienced generations are familiar with only the (semi-)automated process;
    b. Companies may rely heavily on technology experts / solution vendors to develop these algorithms.

  Therefore, it is important for companies to:

    - Maintain clearly defined set of process, control and procedure manuals, as well as robust set of requirement, design and solution specifications for the development work;
    - Retain technical capabilities and skills, at least partially, in-house to avoid being held up by any external vendors; and
    - Provide refresher training, as well as to schedule hands-on practice in the manual undertaking of end-to-end processes for employees.

- *Bias in decision making*: Humans have biases in our decision-making process. However, as some human workers may overestimate a quantity of interest whilst the others may underestimate the same quantity of interest, the concurrent biases observed can be offset against each other to neutralise the overall distortion effect. However, if a singular algorithm decides all cases without the effect of bias cancellation, then there is a risk that even small biases can create a portfolio wide impact. Consequently, it is important to proactively assess algorithmic biases, and allow for validation and finetuning of its workings and performance, and adherence to the defined ethical standards, especially if humans are to be kept out of the loop. In financial services companies, it is common to observe that AI/ML are conservatively being applied to enhance / support human-led decisions, and not (yet) to replace them; As per our use case, our exposure to bias risks is low as the AI/ML is used to support automation of low-risk / low materiality backend processes to enhance efficiency as we review the reliability of the solution, and will refer to manual review if confidence of performance is deemed to be below a defined threshold.

6.2 Strength and weakness of the machine learning solution, and envisaged research and development areas

This project has demonstrated to Zurich Singapore not only the strength and the greater potential in the adoption of ML technologies, but also the expected and unexpected challenges with its implementation. The strength of the transfer learning enabled ML solution for case classification came through when the algorithm can achieve high accuracy after running a training dataset of only 150 examples per category; Similarly, the algorithm can also correctly perform complex entity extraction tasks such as finding company names and dates from complex / large documents with only a few samples for model training. This has confirmed the feasibility in using pre-trained and pre-calibrated models, and thereby making the technology more accessible for use by the Commercial Insurance industry.

This said, there is need to take into consideration the variations in data distributions and data types when preparing training data for ML. Here, the handling of small categories, such as those with fewer than 50 examples per categories, would need to be addressed by way of oversampling or related methods like under-sampling, changes to the model architecture etc., as there are clear limitations and minimum threshold for transfer learning or for that matter any machine learning algorithms to work adequately. As noted in the previous section, whilst high degree of accuracy can be achieved for single label cases, the correct and full identification of multi-label cases remain an ongoing challenge to allow for true and high straight through processing rate to be achieved without human intervention for both single and multi-label cases.

As we reflect on the challenges and potential areas for improvement with the existing solution, it is important that we stay informed in the breakthrough of research and development, and continue our review and experimentation in the following areas:

- New and evolving models and algorithms;
- Approach to address multi-label problems;
- Approach to manage imbalanced data.

With research and development (R&D) actively being pursued by both the Technology industry and the Academics, it is natural to expect the evolution and breakthrough of AI/ML to continue. Latest publications have shown that these new models are 1000 times more powerful than Google BERT, and can understand not only text, but also information in different formats, different languages, etc., and can perform many other tasks such as image classification[25]. Whilst these new approaches are still in an early R&D stage, their application to industry use cases is expected to change over time and will provide an interesting avenue to explore in the future. Further, advances in the distillation of models, i.e. the ability to shrinking the models to improve run time without compromising the performance, would also be eagerly monitored by the team.

To address the multi-label and imbalanced data problems, it is important for the team to

[25] *https://blog.google/products/search/introducing-mum/*

continue its assessment and testing of the relevant model configurations to multi-label case detection (e.g. to direct the model's attention to focus on the handling of multi-label cases[26]) and finding new approaches to manage the skewed dataset without requiring extra data samples to be prepared.

6.3 Roadmap for 2021 and beyond

Whilst the workings of the AI/ML have been demonstrated, more work is required to ensure that the end-to-end process is further fine-tuned and seamlessly embedded into the day-to-day workings of the employees. As the solution is deployed into Production, it is envisaged that re-training of the ML model (i.e. leveraging the feedback data captured at the database as the exceptions are cleared manually by the end users) will take place after 6 – 9 months with aim to improve the performance and accuracy of the underlying model.

In the medium term, it is envisaged that Zurich Singapore will look to consider and adopt additional AI/ML enabled use cases and modules by leveraging the skills and experience, which the Company has accumulated over this POC. On reflection, this project has helped improved our capabilities in the following areas:

- Enhanced the AI-literacy among not only the decision makers but also the operating team;
- Improved our ability to assess the suitability of our use case for AI/ML implementation, and support the go forward implementation of comparable projects (e.g. data gathering, labelling, performance review, testing, etc.);
- Established the foundational technical and data infrastructure required to facilitate and foster future development of AI/ML.

Overall, this puts Zurich Singapore in a good place to further embrace digital opportunities and better realise Zurich's strategy to simplify our way of working, innovate, and deliver superior experience to our customers at an ever-faster pace.

[26] *https://www.kaggle.com/thedrcat/focal-multilabel-loss-in-pytorch-explained*

Appendix A

Overview of the approach adopted by Zurich Singapore that can be considered by other AI/ML practitioners to conduct similar experiment can be found in this Appendix.

A.1 Team setup and project governance

A cross-functional team with specialists from various domains was gathered to implement this highly complex project. The team is comprised of the following members:

- *Commercial Insurance subject matter experts*: These are experienced individuals with significant business and operational knowledge who are familiar with the existing processes and exception handling procedures. They are instrumental to provide the detailed requirement for solution design, extraction of the relevant cases and coordination with onshore / offshore resources in the undertaking of the data mapping exercise for ML, not to mention the user acceptance testing, and provision of training to those impacted by the solution.

- *IT solution architects, service managers and developers*: This include Solution Architects who are responsible for the overall design and documentation of the end-to-end routines and data flows to deliver the expected business requirement; Developers who are responsible for the programming of the APIs to enable connectivity between upstream and downstream components / systems, and to safeguard the orchestration of the end-to-end processes in a robust, fast and coherent manner; and the Service Managers who are responsible for packaging the build to allowing the service to be provided in a stable and supported manner.

- *Project managers and business analysts*: The Project Managers have the overall responsibility to coordinate the work between the different expert teams and to also ensure that the overall project stays on time and budget. The Business analysts are key to ensure that business inputs are correctly "translated" into technical specifications that could be passed on to development and data teams.

- *Data science and machine learning Engineers*: The Data and ML Experts are tasked to setup the overall ML models and data pipelines to enable the automation of the end-to-end process. Most of the work of these experts are devoted to the (pre-)training of the ML models on as many Commercial Insurance data sets as possible, and to the testing of the model to ensure that the received emails with multiple attachments containing complex instructions can be processed successfully.

Weekly meetings, deep-dive workshops (e.g. process reviews), joint working sessions (e.g. to review model outputs), etc. are arranged to enable the coordination of work between the team members.

The project is governed by a Steering Committee that meet on a monthly basis to review progress, approve key decisions / recommendations from the project team, and provide point of escalation to address any issues / risks.

A.2 Project phasing and deliverables

This project is broadly comprised of the following four phases:

- Phase 1: Requirement definition;
- Phase 2: Design and build, for 2A) Data science, and 2B) Integration;
- Phase 3: Testing;
- Phase 4: Operational readiness and go live.

Key activities and deliverables for each of the above outlined phases are set out as follows:

- *Phase 1: Requirement definition* – Under this phase of work:

    - *Understanding of the as-is process*: Key output here is to analyse and define the as-is process by way of holding process walkthroughs as well as hosting live case shadowing sessions, and for the documented process to be validated by the business experts.

    - *Definition of the to-be process*: Key output here is to design and agree the to-be process flow by taking into consideration the preferences from the end users whilst taking note of the constraints from the underlying CRM application where the results would be stored and displayed.

  All in all, a holistic requirement document that outlines the current process (including scope, data flow, exceptions, etc.) as well as the envisaged future process flow is expected to be prepared and signed off by the business – This will then be used as the input to the model development work under Phase 2A, and solution architecture design work under Phase 2B.

- *Phase 2A: Data science design and build* – The scope of work here can be categorised into two areas:

    - *Developing and fine turning of the model*: To begin, the prevailing open-source ML models and architectures should be evaluated based on criteria such as applicability to the business case, availability of the supporting documentation, experience of the team, infrastructure requirement / constraints, etc. – For the purpose of this project, Google BERT was selected for the case classification work while Google ALBERT was adopted for the entity extraction work. Although these models have already been pre-trained using tens of thousands of public documents, it may benefit from additional pre-training, for example for this project, using:
        - Internal documents, to enable the model to identify insurance terms and language more effectively;
        - Labelled data, preferably from the target business process, and covering data from different business quarters to rule out any kind of seasonality effect in the results.

- *Applying additional business logic and configuration to the model*: As the final step of the build, consideration should also be given on whether the model may benefit from the introduction of business rules and supplementary data, to facilitate the definition of its operating boundaries and orchestration of the relevant process steps that cannot necessarily be trained.

- *Phase 2B: Integration design and build* – Two main activities covered here include:
  - *Architecture design and setup*: Taking into consideration the captured business requirement, data flow analysis can then be performed to enable the design for process orchestration and use of the APIs to be defined and developed. The scope of work here should include the setup of the necessary infrastructure / environment to enable the end-to-end process to run in a secured, fast, robust and resilient fashion.

  - *Coding, installation and integration work*: Once the design is defined, build and installation of the various components of the solution can then proceed. For this project, connection points were established between the Outlook inbox and the AI/ML solution, as well as between AI/ML solution and the CRM application.

- *Phase 3: Testing* – After the design, build, installation and setup of the required components, the planned testing activities can then be performed. This is expected to include both IT testing (including unit testing, integration testing etc.) and also business testing (i.e. user acceptance testing, operational readiness verification, etc.) with defined entry and exit criteria. To enable the later, test case examples are extracted / prepared such that as-live review of the workings of the model and the end-to-end solution's performance can be validated by the experienced practitioner of the existing process, and for the identified critical defects to be corrected, before signoff is provided.

- *Phase 4: Operational readiness and go-live* – The business onboarding and support activities covered the following three areas:

  - *Change management*: This is expected to include not only end user trainings, but also several waves of internal communication to ensure that the impacted business users are notified, aware and prepared for the agreed process changes. Further, standard operating procedures (SOP) are to be prepared to provide the step-by-step / screen-by-screen instruction on how the new process is expected to work with a special focus on the exception handling steps that must be performed by the end users, say for the "Amber" or "Red" cases. This is a particularly important aspect of the project i.e. to introduce not only a new technology but empower the end users to master the workings of the new process flow and enabling capability.

- *Post implementation support*: Here, support resources are expected to be secured to ensure that the relevant technical subject matter experts are available to swiftly resolve any post go-live issues before the service is handed over to the relevant IT service management team. In addition, it is important for “Change champions” from the existing processing team to be identified and trained, to operate as the first point of contact for internal staff, should they have any questions and concerns.

- *Documentation*: To ensure the necessary knowledge and knowhow is retained, there is need to ensure that the relevant project documentation, including but not limited to the architecture / dataflow diagrams, process flow / procedure manuals, training document, etc. are prepared, packaged and signed off.

After the completion of the project, post implementation review is to be performed and feedback from the users is to be gathered to ensure that learnings from the project can be captured, and ideas for process / system / modelling and improvements can be systematically gathered and evaluated, where applicable, be prioritised and considered for implementation in the subsequent release of the solution.