IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Supervisory letter writing app: expediting letter drafting and ensuring tone consistency[1]

Joshua Tan, Chi Ken Shum and Mohd Akmal Amri,
Central Bank of Malaysia

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Supervisory Letter Writing App: Expediting Letter Drafting and Ensuring Tone Consistency[1]

Joshua Tan[2], Chi Ken Shum[3] and Mohd Akmal Amri[4]

Abstract

The Central Bank of Malaysia issues supervisory letters to all financial institutions under its purview on an annual basis. The supervisory letters communicate the supervisory ratings assigned to the respective institutions, the supervisory concerns as well as the remediation actions required. With more than 120 institutions under purview, writing and reviewing every supervisory letter consumes hundreds of costly work hours. In an effort to improve both the efficiency in the writing process as well as the consistency in the communication tone of supervisory letters, we develop a web application that comprises two main features: i) Tone Analysis; ii) Sentence Search. With Tone Analysis, we empower supervisors to better understand and calibrate the tone of their drafted letters commensurate with the intended corrective measures to be taken by financial institutions. We utilize a Transformer-based Natural Language Processing (NLP) model called DistilBERT coupled with optimizations using ONNX Runtime and quantization to perform multi-class text classification on every sentence in a letter. Each sentence is classified into one of four ordinal tone classes — "Neutral", "Cautious", "Concerned", and "Forceful" with tone ranging from no tone ("Neutral") to the most severe tone ("Forceful"). Using the weighted-F1 metric, we obtain a score of 77.6% for our test data set. With Sentence Search, we enable supervisors to search for sentences extracted from previously issued letters by tone class to expedite the writing process. Supervisors can search for sentences by keywords or by semantic similarity with the latter utilizing an NLP model called Sentence-BERT.

Keywords: Central bank communication, prudential supervision, supervisory letter, natural language processing, sentiment, deep learning, web application.

JEL classification: C55, C80, E58, G28, Y80.

[1] This paper was prepared for the IFC and Bank of Italy Workshop on "Data Science in Central Banking" held virtually on 19-22 October 2021. The views expressed in this paper are those of the authors and should not be interpreted as reflecting the views of the Central Bank of Malaysia or of anyone else associated with the Central Bank of Malaysia. The authors are grateful to Daniel Chin Shen Li, Intan Zainura Ahmad and Syairulzila Mohd Yusof for their assistance and useful comments on the paper.

[2] Bank Negara Malaysia; Email: joshua@bnm.gov.my

[3] Bank Negara Malaysia; Email: shum@bnm.gov.my

[4] Bank Negara Malaysia; Email: akmalamri@bnm.gov.my

Contents

1. Introduction

As part of the Central Bank of Malaysia's mandate to preserve the country's financial stability, all licensed financial institutions (FIs) operating in Malaysia are subject to annual supervisory review by the Central Bank of Malaysia, where the safety and soundness of each FI is assessed and determined by their respective supervisors. The supervisors are guided by a robust risk-based supervisory framework where each FI is evaluated based on its risk profile, quality of risk management, sustainability of earnings, as well as strength of capital and liquidity before being assigned with a composite supervisory rating. At the conclusion of their review, supervisors will issue supervisory letters to the respective FIs to communicate their assigned supervisory ratings, supervisory concerns as well as the remediation actions required.

Currently, there are more than 120 FIs under the central bank's purview and therefore, at minimum 120 supervisory letters need to be issued every year. Each supervisory letter is drafted and reviewed thoroughly by the respective supervisors, and it may take weeks before the letter is fit for issuance as it goes through the necessary editing and governance processes. Based on our engagement with supervisors, most supervisors who are relatively new to the position find drafting supervisory letters laborious due to the numerous revisions that occur during the review process. This stringent review process is necessary due to the subjective nature of adjusting the tone of the supervisory letter when detailing supervisory concerns and the required remediation actions. For example, the Central Bank of Malaysia's management would expect a strongly worded letter when conveying regulatory breaches or recurring lapses in control functions. However, having a good grasp of the appropriate words and sentences to use under specific situations when drafting a supervisory letter is predominantly a result of tacit knowledge gained through years of experience. Therefore, inexperienced supervisors often find this aspect challenging.

In addition, as supervisory letters are issued by multiple supervisory departments, there is no effective nor systematic way to streamline the tones used in these letters across the departments. There are a variety of factors that influence the tones used in the supervisory letters. These include the severity of issues highlighted, their sentiment regarding the competency of personnel to whom they address the issue and their personal writing style. While variety in tone is necessary in some cases to address situations unique to certain institutions, having a tool to streamline the way supervisors write letters will help to reduce the wide range of individual writing styles and adopt a more consistent communication strategy. Such communication tone is critical as supervisory letters reflect the central bank's overall assessment of an FI and its human capital. Thus, achieving a precise tone in the letters will establish clear expectations on the subsequent actions the FIs must undertake.

As such, we develop a web application with two main features to address the challenges mentioned above. The first feature of the web application is Tone Analysis. Tone Analysis enables supervisors and management to gauge the tone employed in a letter and determine if it is proportionate and appropriate to the risk profile and severity of the issues highlighted to the FI. The second feature, Sentence Search, expedites the letter drafting process as supervisors can easily query a database of sentences from historical supervisory letters while drafting letters. When used together, these features enhance the supervisory process by improving supervisors' efficiency in drafting supervisory letters and by maintaining a consistent communication approach for the Central Bank of Malaysia.

2. Related Work

2.1 Central Bank and Supervision Domain

Much effort in analyzing central bank communications using Natural Language Processing (NLP) have revolved around sentiment analysis and topic analysis in the context of public-facing monetary policy and financial stability communications. The sentiments are often summarized into an index that is then studied in relation to a set of economic or financial indicators. Examples of such studies include Correa et al. (2017), Jegadeesh and Wu (2015) and Born et al. (2014). At the same time, there has also been increased interest in the broader application of NLP within the supervision domain. Many supervisory authorities, including central banks, have embarked on studies, projects and proof-of-concepts applying new technologies in the realm of supervision, engendering the frequently dubbed term "SupTech". Financial Stability Board (2020) highlights this fact with the numerous case studies of applied NLP for content extraction, risk identification and news-based sentiment analysis in relation to supervisory matters.

Our project focuses on analyzing the sentiment or tone of central bank communications with regulated FIs — a private but crucial stakeholder. In this respect, Bholat et al. (2017) is one notable example that resembles our initiative as they studied the linguistic features of the letters sent by the Bank of England's Prudential Regulation Authority (PRA) to banks and building societies under its supervision. However, unlike Bholat et al. (2017), this paper does not discuss the characteristics of supervisory letters such as their directiveness and formality in depth but focuses on the methodology applied in developing a SupTech tool to facilitate the letter drafting process.

2.2 Data Science Domain

In our web application, the main techniques we employ are text classification (more specifically, sentiment analysis) and sentence embedding. Text classification is an NLP technique used to categorize a sequence of text into groups. It has made significant progress in recent years with state-of-the-art deep learning models leapfrogging traditional machine learning models in performance (Minaee et al., 2021). This advancement is in part due to the groundbreaking Transformers architecture proposed by Vaswani et al. (2017) coupled with transfer learning techniques proposed by Howard and Ruder (2018). While there has been some effort in utilizing text classification to analyze sentiment in central bank communications (Rybinski, 2019), most of the literature studied dictionary-based approaches as seen in Shapiro and Wilson (2019), Hubert and Labondance (2017) and Fraccaroli et al. (2020).

Meanwhile, in the broader data science domain, text classification is commonly used for sentiment analysis. This is evidenced by the various sentiment-related data sets used as benchmarks for the evaluation of state-of-the-art text classification models. At present, a variety of models have achieved top performance across a broad array of tasks (binary or multiclass classification) and data sets (IMDb, Yelp etc.). Some noteworthy models are XLNet (Yang et al., 2019), Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), Universal Language Model Fine-tuning (ULMFiT) (Howard & Ruder, 2018) along with their variations.

On the other hand, word and sentence embeddings are dense vectors or numeric representations of a text that, when produced using machine learning models, can capture its semantic meaning. These representations are commonly generated as an intermediate step for text classification models. While they are key components for many NLP models, they can also be used standalone to serve other purposes such as computing cosine similarities for the search of similar sentences. Modern embedding models have evolved significantly, with popular models like word2vec (Mikolov et al., 2013) and Global Vectors for Word Representation (GloVe) (Pennington et al., 2014) being able to create static word embeddings that combine all the different meanings of a word into one vector. Nonetheless, newer models like Embedding Language Model (ELMo) (Peters et al., 2018) and BERT have advanced to generate dynamic word embeddings that change its representation according to the context of a sentence, producing superior results in most cases.

3. Web Application

The Django web application we develop has two main features: i) Tone Analysis; and ii) Sentence Search. The technical details of their implementations are provided in the Methodology section below. In this section, we describe their business use cases.

With Tone Analysis, we empower supervisors to better understand and calibrate the tone of their drafted letters commensurate with the intended corrective measures to be taken by FIs. Supervisors can upload a draft letter and obtain a sentence level tone prediction that is color-coded according to tone for easier analysis. For further insight, they can choose to interpret the predicted tone of a sentence which will display the confidence score of the prediction and highlight keywords that contribute most to the prediction. If a supervisor intends to draw attention to a key risk faced by an FI via a strong directive message but receives a soft predicted tone from the model, he or she may want to re-evaluate the language of the sentence. For example, specific words that are deemed to significantly influence the predicted tone may be edited to improve the intended message severity level. Conversely, if the predicted tone is consistent with the supervisor's intention, the application may serve as a form of validation before letter issuance. Supervisors can also view the tone distribution of all predicted sentences summarized in a chart. These sentences classified into their respective tones each contribute a specific score that is aggregated and normalized using a formula to arrive at a compound score for the entire letter. The compound score represents the overall tone of the letter and is assigned to a category according to predefined thresholds for document-level comparisons. Regularly using the application to run tone analyses over the course of drafting supervisory letters ensures that the Central Bank of Malaysia, across its multiple departments, communicates to FIs in a consistent manner.

With Sentence Search, we enable supervisors to search for sentences extracted from previously issued letters, categorized by tone to expedite the writing process. As supervisors draft letters, the need to reference past supervisory letters issued to FIs with similar issues or similar ratings often arises. This reference process is necessary to provide context for the issue at hand to supervisors and to subsequently prevent FIs from overreacting or underreacting to sudden changes in vocabulary, writing style and in turn, communication tone in the supervisory letters. Supervisors can either search by keywords, e.g., "liquidity risk" or by a full sentence. When searching by sentence, the application will extract sentences that are semantically similar to the query sentence even though they may not be identical. For example, given a query “Loan growth has declined due to poor GDP growth”, a possible search result is “Due to worsening economic condition, loan growth has shrunk.”. To perform a search, supervisors simply need to type in the word or sentence of interest in an input box to query the database. An alternative workflow would be the following — after a supervisor has uploaded a letter for tone analysis, he or she may identify specific sentences that do not convey the intended tone and seek to amend them. The supervisor can directly highlight words or sentences from the tone analysis page and click a button to query the database. Once the most similar sentences are presented, supervisors can also expand each search result to read the original letter content for a better understanding of the context in which the sentence was used. Ultimately, these search features enable supervisors to effectively query and reference historical letters, leading to stronger draft letters, thereby reducing the total time spent on amendments as they alternate between author and reviewer.

While these two features aim to improve the letter drafting process for supervisors, we also developed a privileged access dashboard in the web application for administrators. The dashboard provides easy retrieval of past letters and a holistic visualization of overall letter tones alongside simple filtering options for time series and cross-sectional analyses. This view is valuable for monitoring and spotting anomalies in letter tones across FIs of similar risk ratings or within a particular FI across time.

4. Methodology

4.1 Tone Analysis

We perform multi-class text classification on sentences extracted from past supervisory letters into four ordinal tone classes — “Neutral”, “Cautious”, “Concerned” and “Forceful” with tone severity ranging from none (“Neutral”) to the most severe (“Forceful”). We choose the weighted-F1 score as our metric due to an imbalanced class distribution and the relatively equal importance of precision and recall for our use case.

4.1.1 Data Preprocessing and Labeling

Our data set is derived from a collection of confidential supervisory letters issued over a period of four years between 2013 and 2016. The letters are first processed into 15,000 individual sentences using a combination of regular expression and custom rules, then anonymized using a custom Named-entity recognition (NER) model. The NER model identifies sensitive information such as financial institution names, corporation names, and individual names that are then masked for data security reasons. Further, these anonymized entities may prevent the trained model from associating certain words with certain financial institutions, potentially reducing model bias.

Once data preprocessing is complete, the sentences are manually labeled as one of the four aforementioned tone classes by experienced supervisors. This serves as training data for our text classification model and as ground truth for model evaluation. Every sentence is labeled independently by three supervisors according to a labeling guide with the majority vote for each sentence selected as the final label. In addition, the labeled data is reviewed by a fourth supervisor on a sampling basis as quality check to ensure consistency of the labels. The entire labeling process is divided into six separate rounds, with supervisors reconvening after each round to discuss difficulties that they encounter, especially for sentences that are labeled differently by all three supervisors. Further, the periodic recalibrations also provide the opportunity for supervisors to jointly update the labeling guide with new insights, retroactively amend labeling from earlier rounds and evaluate intermediate model performance.

In our context of supervisory communication, labeling for sentence tone is challenging due to the nuanced language used in supervisory letters. Generally, supervisors need to identify parts of the sentence that best reflect the intended tone without accounting for the severity of the supervisory issue at hand. For example, in the sentence "inadequate details presented in credit risk reports", the word "inadequate" should convey tone, but not the credit risk report reference. This approach is necessary as the range of issues that concern an FI can vary tremendously. Therefore, if all possible types of issues are taken into consideration during labeling, the machine learning model would likely underperform considering the multiple tone classes and small data size. Nonetheless, there are exceptions provided for specific issues like money laundering and terrorist financing risks that are taken very seriously by the Central Bank of Malaysia. These sentences warrant a deviation in tone classification due to the severity of the issue.

4.1.2 Model and Performance

As mentioned earlier, deep learning models have been making breakthrough performances in text classification. However, these models are typically large and require significant computing power for training and inferencing. Due to compute limitations, we restrict our model search to smaller and lighter models that are able to train quickly and perform inference effectively. Hence, we choose DistilBERT (Sanh et al., 2020), a smaller and faster version of the original BERT model proposed by Google, that retains much of BERT's performance with a significantly lower number of parameters (66 million vs 110 million).

Similar to BERT, DistilBERT employs transfer learning, a technique of pre-training a model on vast amounts of general-domain internet text data before transferring that knowledge to a downstream task. This process gives the language model a significant boost in performance as it learns a diverse range of word usage in various linguistic structures. We first download the pre-trained DistilBERT model provided by Hugging Face. However, instead of directly finetuning DistilBERT on a downstream task like text classification, we implement the target task language model fine-tuning technique proposed by Howard and Ruder (2018) for ULMFiT and demonstrated by Sun et al. (2019) for BERT. This entails additional pre-training on both the training and test data set to update the parameters in DistilBERT's masked language model in order to better reflect information from a prudential supervision domain. This is also necessary since supervisory letters may contain words that are only used in a local context, such as those derived from Malaysian regulatory requirements. We then integrate the fastai library (Howard & Gugger, 2020) with Hugging Face's transformers library (Wolf et al., 2020) as shown by Roberti (2019) to leverage on discriminative fine-tuning, slanted triangular learning rates and gradual unfreezing for training after attaching the classifier layer. These additional features provided by the fastai library stem from an Idea derived from the sub-field of computer vision, whereby earlier layers of a neural network contain more general features, while later layers learn features specific to the task (Yosinski et al., 2014). Therefore, we use these techniques to achieve granular control over the training extent for each layer — later layers of DistilBERT are trained more than earlier layers for better classification ability. Due to the small data size, we limit the number of training epochs to a small number to prevent the model from overfitting. We find that overall, this additional fine-tuning step improves the weighted-F1 score by 1.4%, resulting in a final weighted-F1 score of 77.6% as shown in Table 1.

Model performance on test set

In percentage (%) Table 1

Model	Accuracy	F1 (weighted)
Bag-of-Words + Logistic Regression	66.0	65.4
Bag-of-Words + XGBoost	67.0	65.3
SBERT + Logistic Regression	70.2	69.4
SBERT + XGBoost	69.1	67.7
DistilBERT-FiT	77.8	77.3
DistilBERT-FiT (Quantized)	74.6	73.0
DistilBERT-FiT + ONNX Runtime (Quantized)	76.8	76.2
DistilBERT-ITPT-FiT	**78.8**	**78.4**
DistilBERT-ITPT-FiT (Quantized)	74.8	74.4
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	78.0	77.6

Note: SBERT is "Sentence-BERT". DistilBERT-FiT is "DistilBERT + downstream FineTuning". DistilBERT-ITPT-FiT is "DistilBERT +8nto8sn-Task Pre-Training + downstream Fine-Tuning". Our final selected model is DistilBERT-ITPT-FiT + ONNX Runtime (Quantized).

At the initial project exploration stage, we applied a simple lexicon-based approach by identifying and grouping domain-specific words that are commonly used in supervisory letters. Each of these words were given a tone score according to its group to indicate tone severity which could then be used to automatically analyze and score a full-length supervisory letter. Not surprisingly, this approach proved difficult to account for the various word forms, combinations, semantics, and grammatical structure within the letter. For example, if a section in a letter contained one word with a severe tone but many words with softer tones, its overall tone could be ambiguous. Given that the web application serves to aid supervisors in validating an intended tone or recalibrating an unintended tone, this situation may be difficult for supervisors to interpret.

Hence, we decided to proceed with supervised machine learning where we performed multi-class text classification on every sentence in a letter. Despite the loss of some tonal information when a text sequence is evaluated on a sentence level without full context from the original paragraph, the trade-off for an overall simpler and more interpretable tone was worthwhile. We first experimented with computationally cheaper models, starting with a Bag-of-Words feature representation for sentences which were then piped in as input to various traditional machine learning models like logistic regression and Extreme Gradient Boosting (XGBoost) (Chen & Guestrin, 2016). While this overcame some of the issues mentioned earlier, these vectors still lacked semantic meaning and contextual information. Thus, we quickly progressed to a Transformer-based NLP model called

Sentence-BERT (SBERT). SBERT is a modification of the pre-trained BERT model that uses9nto9see and triplet network structures to derive dense vectors that can capture semantic meaning and contextual information of words in a sentence (Reimers & Gurevych, 2019). These generated sentence embeddings were then fed into a downstream machine learning classification model.

Despite yielding performance gain, the sentence embeddings were created from a model trained on a general-domain text corpus which may not have been suitable for the prudential supervision domain. In order to achieve better performance, we required a model embedded with a classifier which could update its weights for a text classification task in the prudential supervision domain, leading to our final selected model — DistilBERT.

4.1.3 Model Inference Optimizations

Inference time per sentence

In milliseconds (ms)

Table 2

Framework	Min	Mean	Median	Max
DistilBERT-ITPT-FiT	13.78	30.70	28.67	96.57
DistilBERT-ITPT-FiT + ONNX Runtime	6.59	20.52	19.65	173.64
DistilBERT-ITPT-FiT (Quantized)	5.85	18.70	17.02	145.49
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	**2.67**	**12.69**	**11.89**	**52.70**

Inference time per letter

In seconds (s)

Table 3

Framework	Min	Mean	Median	Max
DistilBERT-ITPT-FiT	0.43	2.98	2.42	9.16
DistilBERT-ITPT-FiT + ONNX Runtime	0.28	1.99	1.62	5.78
DistilBERT-ITPT-FiT (Quantized)	0.23	1.81	1.38	5.28
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	**0.17**	**1.23**	**1.02**	**3.64**

Note: Distributions of inference times for a sentence and a letter are presented in Figure 1 and Figure 2 respectively in Appendix I.

To optimize inference (text classification) speed, we use ONNX Runtime and quantization on DistilBERT to reduce the median inference time for a sentence by 2.4 times from 28.67 milliseconds to 11.89 milliseconds with minimal impact to the weighted-F1 score (decreased by 0.8%). ONNX Runtime is an optimization library that boosts inference speed while quantization is a technique that approximates floating-point numbers with integers to reduce memory use and accelerate performance (Functowicz and Li, 2020). As supervisors typically need to repeatedly upload different versions of their letters when drafting, it is imperative that we maintain acceptable inference time for a full-length letter while maximizing the weighted-F1 metric. Even though a baseline DistilBERT without optimizations ("DistilBERT-ITPT-FiT") would perform inferencing faster than a larger baseline BERT, limited computational resources would still dampen user experience of the web application as users could potentially wait up to 9.16 seconds for a letter to be analyzed. This necessitates the use of ONNX Runtime and quantization optimizations, performing inference on a letter with a median of 1.02 seconds and a maximum of 3.64 seconds.

4.1.4 Model Prediction Interpretation

In order to interpret a predicted sentence tone, we utilize integrated gradients (Sundararajan et al., 2017) implemented in Captum (Kokhlikyan et al., 2020), a model interpretability library for PyTorch. We also display the probability of the model's prediction when presenting the integrated gradients output. Integrated gradients is a method that attributes a model's prediction on an input to features of the input. Th10ntouition is the following — integrated gradients generates a linear interpolation from a baseline vector to a final vector that represents our predicted sentence, effectively producing multiple vectors that are progressively closer to the final vector. It then acquires gradients which significantly influence the prediction probability across the interpolated vectors and averages them. For our use case, integrated gradients is able to attribute the predicted tone of a sentence to every word in it with varying contribution levels. Words with positive attribution scores nudge the sentence towards the predicted tone while those with negative attribution scores pull the sentence away from it. In practice, this often results in specific keywords visibly highlighted for their positive attributions, reflecting the specific vocabulary used by supervisors when communicating issues of different severities (Appendix II). For example, in the sentence "inadequate compliance review and limited audit coverage scope", integrated gradients may highlight the words "inadequate" and "limited" as keywords that contribute most to the predicted sentence tone. When paired with the prediction probability score, supervisors can assess the reliability of this prediction and interpretation wherein a lower probability indicates greater model prediction uncertainty and therefore may not be relied on.

4.1.5 Machine Learning Workflow

To ensure a sustainable machine learning lifecycle that continuously produces relevant results, we incorporate a machine learning workflow component to support intuitive model training, tracking and deployment. This is accomplished in part via the integration of MLflow (Zaharia et al., 2018), a tool that greatly simplifies these operations. The simplicity of MLflow allows us to extend its features to enable supervisory administrators to manage the web application without significant input from the technical team.

The machine learning workflow begins with supervisors uploading finalized supervisory letters that are ready for storage into the web application. The application preprocesses the letters into a list of sentences, masks the sensitive information using a custom NER model, then stores them in Elasticsearch. Thereafter, the data labeling team can generate the list of preprocessed sentences for labeling. Once the sentences are labeled, the file is uploaded back into the web application. This action triggers a model training run that produces a new version of the model alongside performance metrics and hyperparameter configurations. Supervisors can then easily compare the model's performance across different training runs and select the best model for deployment.

In addition to the periodic data labeling by a dedicated team, we supplement the labeled dataset with input from the community of supervisors who utilize the Tone Analysis feature to analyze their draft letters. As they analyze the tone of their draft supervisory letters, supervisors can provide input on the model's predicted sentence tone by suggesting a different tone classification via the user interface. These user feedbacks are then aggregated and reviewed by an administrator for potential inclusion in the next cycle of model retraining.

4.2 Sentence Search

Sentence Search enables search by keywords or by semantically similar sentences extracted from historical supervisory letters stored in Elasticsearch, a database optimized for text queries. With keyword matching, all sentences that contain the query will be extracted with options to filter search results by tone.

To implement sentence search by semantic similarity, we utilize a pre-trained SBERT ("all-mpnet-base-v2"), fine-tuned from Microsoft's MPNet model (Song et al., 2020). Similar to DistilBERT, SBERT can provide semantically meaningful sentence embeddings, but does not include a classifier layer required for text classification and is optimized for tasks like semantic textual similarity. We attempted to further train the SBERT model using existing supervisory sentences to adapt it for the supervision domain, but the fine-tuned model performed worse than the original model due to limited data (insufficient examples of highly similar sentences).

As a result, we directly use the pre-trained model to encode the sentences from all historical supervisory letters into sentence embeddings that we then store in Elasticsearch alongside the original sentences. When a supervisor inputs a query sentence through the web application, the query is converted to an embedding using the same SBERT model. With both the query and historical sentences embedded in the same vector space, we can calculate the cosine similarity scores to determine how similar the sentences are. Subsequently, we rank the sentences by similarity scores and present the relevant results. Once again, the search results can then be filtered by tone.

5. Conclusion

In this project, we develop a web application to facilitate the drafting of supervisory letters and to ensure consistency in their communication tone. We accomplish this via text classification and semantic textual similarity in the tone analysis and sentence search features respectively. For future work, we seek to re-evaluate the labeled data to better distinguish the tone classes with the aim of improving the weighted-F1 score. This may be performed in parallel with the labeling of additional data derived from more recent supervisory letters to update the model for changes in language and supervisory issues. Further, we intend to develop additional tools surrounding supervisory issues and remediation actions detailed in the letters that may be integrated with this Supervisory Letter Writing App to form a holistic web application.

References

Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). Text mining for central banks. *Available at SSRN 2624811*.

Bholat, D., Brookes, J., Cai, C., Grundy, K., & Lund, J. (2017). Sending Firm Messages: Text Mining Letters from PRA Supervisors to Banks and Building Societies They Regulate. Bank of England Working Paper, 688. https://doi.org/10.2139/ssrn.3066809

Born, B., Ehrmann, M., & Fratzscher, M. (2014). Central Bank Communication on Financial Stability. The Economic Journal, 124(577), 701–734. https://doi.org/10.1111/ecoj.12039

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016, 785–794. https://doi.org/10.1145/2939672.2939785

Correa, R., Garud, K., Londono, J. M., & Mislang, N. (2017). Sentiment in Central Bank's Financial Stability Reports. International Finance Discussion Paper, 1203, 1–46. https://doi.org/10.17016/ifdp.2017.1203

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171–4186. https://10.18653/v1/N19-1423

Fraccaroli, N., Giovannini, A., & Jamet, J.-F. (2020). Central banks in parliaments: a text analysis of the parliamentary hearings of the Bank of England, the European Central Bank and the Federal Reserve. ECB Working Paper, 20202442.

FSB. (2020). The Use of Supervisory and Regulatory Technology by Authorities and Regulated Institutions: Market developments and financial stability implications. https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2442~e78be127c0.en.pdf

Howard, J., & Gugger, S. (2020). Fastai: A Layered API for Deep Learning. Information, 11(2), 108. https://doi.org/10.3390/info11020108

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 1, 328–339. https://doi.org/10.18653/v1/p18-1031

Hubert, P., & Labondance, F. (2018). Central Bank Sentiment and Policy Expectations. Bank of Englang Working Paper, 648. https://doi.org/10.2139/ssrn.2920496

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., & Reblitz-richardson, O. (2020). Captum : A unified and generic model interpretability library for PyTorch An Overview of the Algorithms. ArXiv, abs/2009.07896

Li, Y. (2020, September 1). Faster and smaller quantized NLP with Hugging Face and ONNX Runtime. Medium. https://medium.com/microsoftazure/ faster-and-smaller-quantized-nlp-with-hugging-face-and-onnx-runtime-ec5525473bb7

Jegadeesh, N., & Wu, D. (2015). Deciphering Fedspeak: The Information Content of FOMC Minutes. Working Paper, University of Pennsylvania.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations (ICLR), 1–12.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep Learning Based Text Classification: A Comprehensive Review. ArXiv, abs/2004.03705

Pennington, P., Socher, R., & Manning C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532– 1543. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 2227–2237. https://doi.org/10.18653/v1/n18-1202

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982–3992. https://doi.org/10.18653/v1/d19-1410

Roberti, M. (2019, November 27). Fastai with Transformers (BERT, RoBERTa, XLNet, XLM, DistilBERT). Towards Data Science. https://towardsdatascience.com/fastai-with-transformers-bert-roberta-xlnet-xlm-distilbert-4f41ee18ecb2

Rybinski, K. (2019). A machine learning framework for automated analysis of central bank communication and media discourse. The case of Narodowy Bank Polski. Bank i Kredyt, 50(1), 1–19.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108

Shapiro, A. H., & Wilson, D. (2019). Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives using Text Analysis. Federal Reserve Bank of San Francisco Working Paper, 2. https://doi.org/10.24148/wp2019-02

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. NeurIPS, 1–14. https://arxiv.org/abs/2004.09297

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? Lecture Notes in Computer Science, 11856, 194–206. https://doi.org/10.1007/978-3-030-32381-3\16

Sundararajan, M., Taly, A., & Yan, Qiqi. (2017). Axiomatic Attribution for Deep Networks. ArXiv, abs/1703.01365v2

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. Advances in Neural Information Processing Systems, 6000-6010.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). Transformers: State-of-the-art natural language processing. Proceedings of the 2020 Conference on Empirical Methods in

Natural Language Processing: System Demonstrations, 38-45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

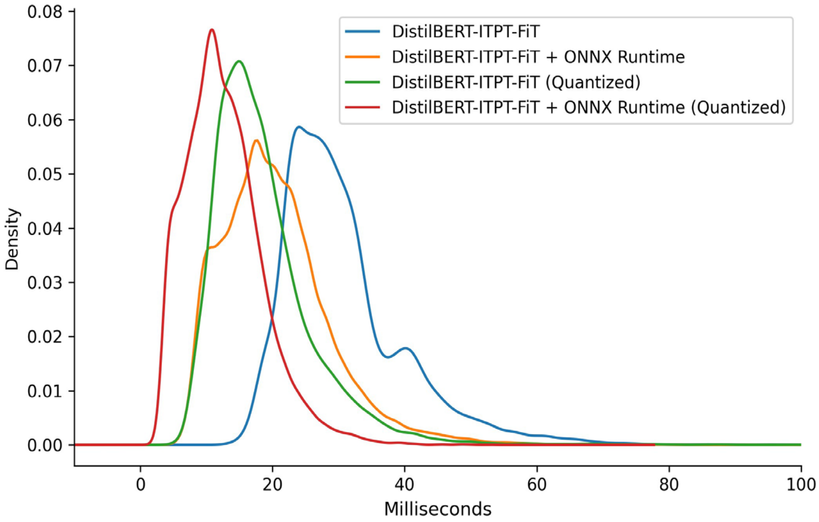
Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Aautoregressive Pretraining for Language Understanding. NeurIPS.

Zaharia, M., Chen, A., Davidson, A., Ghodsi, A., Hong, S. A., Konwinski, A., ... & Zumar, C. (2018). Accelerating the machine learning lifecycle with MLflow. *IEEE Data Eng. Bull.*, *41*(4), 39-45.

Appendix I: Inference Time Distribution Graphs

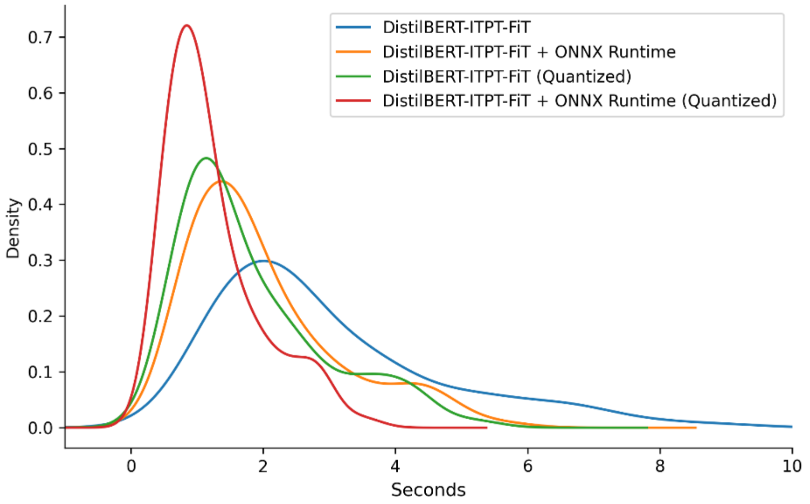
Distribution of inference time for a sentence

Graph 1



Distribution of inference time for a letter

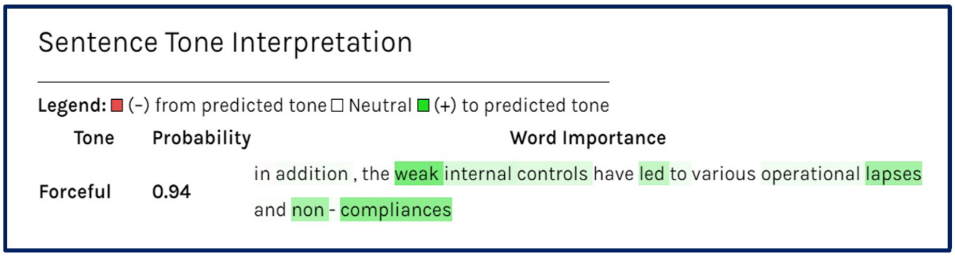
Graph 2



Appendix II: Model Prediction Interpretation

Interpretation for a sentence with predicted “forceful” tone

Figure 1



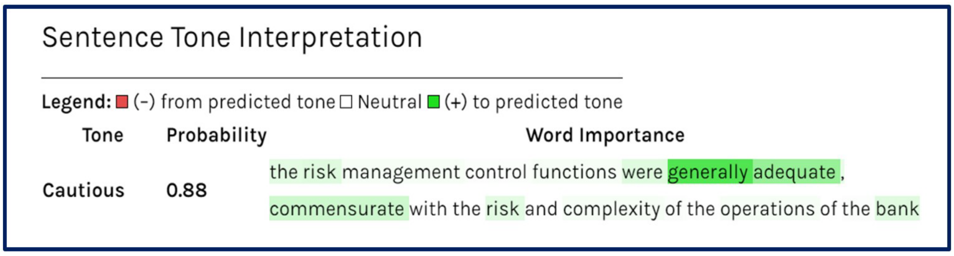
Sentence Tone Interpretation

Legend: 🟥 (–) from predicted tone □ Neutral 🟩 (+) to predicted tone

Tone	Probability	Word Importance
Forceful	0.94	in addition , the weak internal controls have led to various operational lapses and non - compliances

Interpretation for a sentence with predicted “cautious” tone

Figure 2



Sentence Tone Interpretation

Legend: 🟥 (–) from predicted tone □ Neutral 🟩 (+) to predicted tone

Tone	Probability	Word Importance
Cautious	0.88	the risk management control functions were generally adequate , commensurate with the risk and complexity of the operations of the bank

Supervisory Letter Writing App

Expediting Letter Drafting & Ensuring Tone Consistency

IFC-Bank of Italy Workshop on Data Science in Central Banking

Authors: Joshua Tan, ***Chi Ken Shum****, Mohd. Akmal*

Background

Background

- Financial institutions (FIs) operating in Malaysia are subject to annual supervisory examinations by Bank Negara Malaysia, where the safety and soundness of each FI is assessed.
- At the conclusion of their review, supervisors will issue supervisory letters to the respective FIs to communicate their assigned supervisory rating, highlight supervisory concerns and recommend remediation actions required.

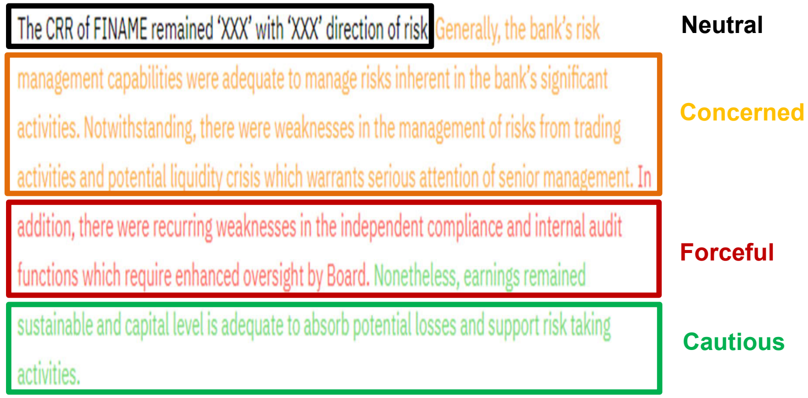
Problem Statement

- Most supervisors who are relatively new to the position find drafting supervisory letters is laborious due to the numerous revisions that occur during the review process.
- Supervisory letters are issued by multiple departments. Hence, there is no effective nor systematic way to gauge the tones used in these letters.

Web Application (1)

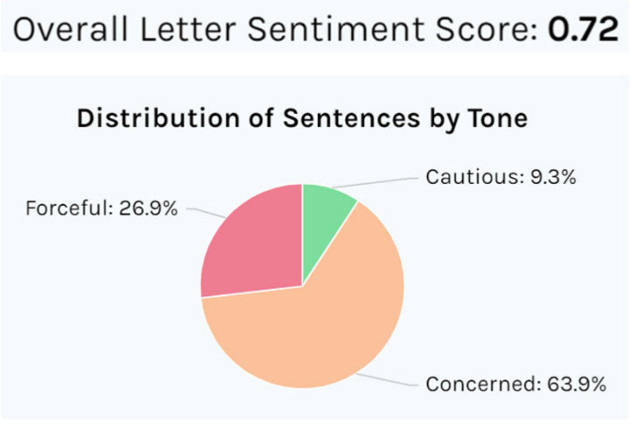
Module 1: Tone Analysis

- Facilitate supervisors to better understand and calibrate the tone of their drafted letters to commensurate with the supervisory concerns and intended corrective measures to be taken by FIs.
- Supervisors can upload a draft letter and obtain sentence level tone predictions that are color-coded according to tone for easier analysis.
- Supervisors can also view the overall letter sentiment score and the tone distribution of all predicted sentences summarized in a chart.



The CRR of FINAME remained 'XXX' with 'XXX' direction of risk Generally, the bank's risk management capabilities were adequate to manage risks inherent in the bank's significant activities. Notwithstanding, there were weaknesses in the management of risks from trading activities and potential liquidity crisis which warrants serious attention of senior management. In addition, there were recurring weaknesses in the independent compliance and internal audit functions which require enhanced oversight by Board. Nonetheless, earnings remained sustainable and capital level is adequate to absorb potential losses and support risk taking activities.

Supervisory letter with predicted sentence tones



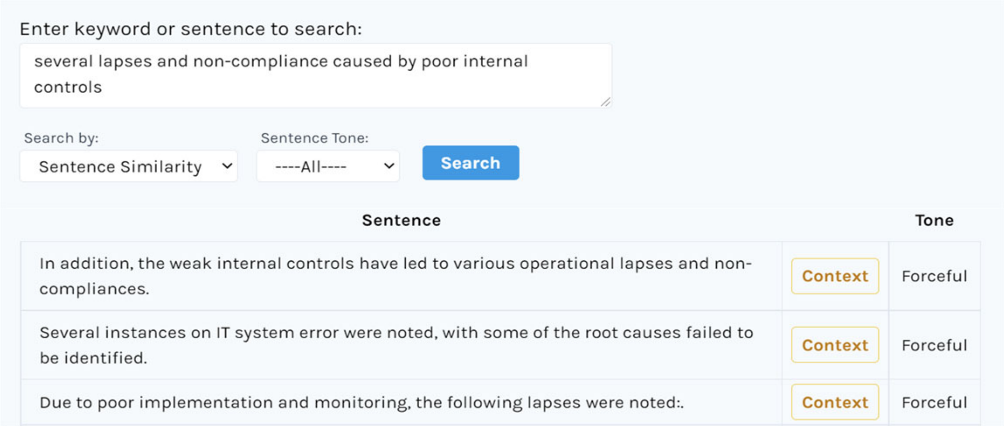
Overall letter score and sentence tone distribution

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

Web Application (2)

Module 2: Sentence Search

- Enable supervisors to search for sentences extracted from previously issued letters by tone to expedite the writing process.
  - – Important for context understanding, consistency in vocabulary.
- How to perform a search?
  1. i. Navigate to search page → Type query in input box
  2. ii. After uploading the supervisory letter for tone analysis → Highlight word/ sentence of interest and query the database directly from the tone analysis page
- Can search by keywords or semantically similar sentences.



Enter keyword or sentence to search:

several lapses and non-compliance caused by poor internal controls

Search by: Sentence Similarity

Sentence Tone: ----All----

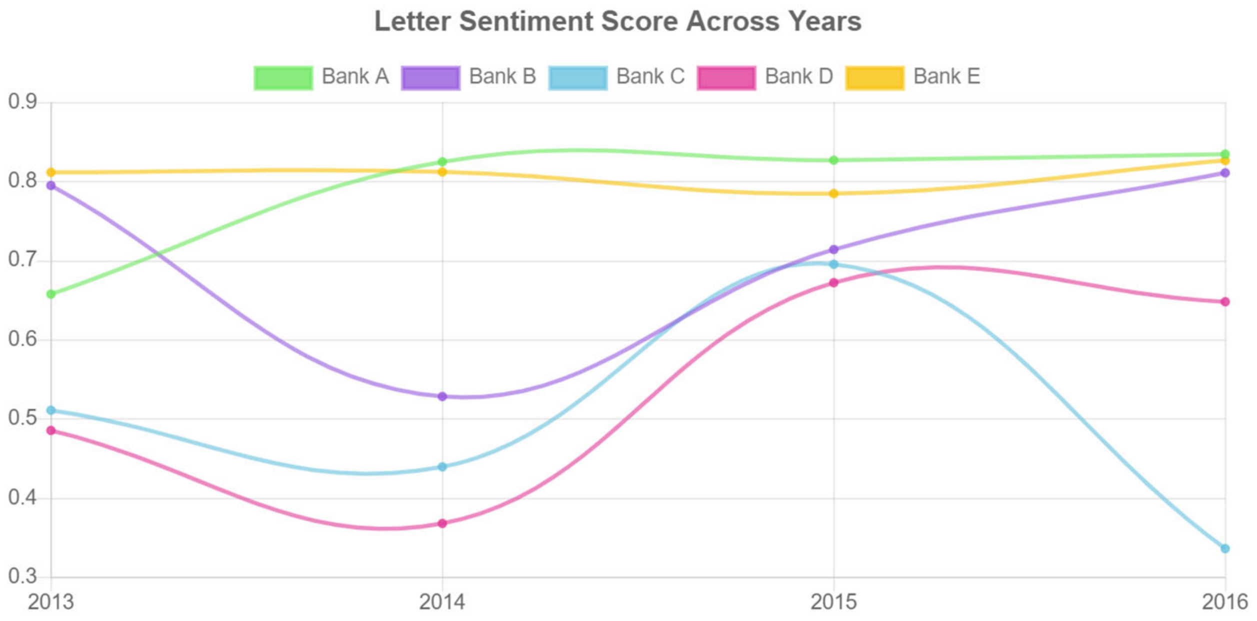
Sentence		Tone
In addition, the weak internal controls have led to various operational lapses and non-compliances.	Context	Forceful
Several instances on IT system error were noted, with some of the root causes failed to be identified.	Context	Forceful
Due to poor implementation and monitoring, the following lapses were noted:.	Context	Forceful

Search result of a sentence by semantic similarity

Web Application (3)

Admin Dashboard

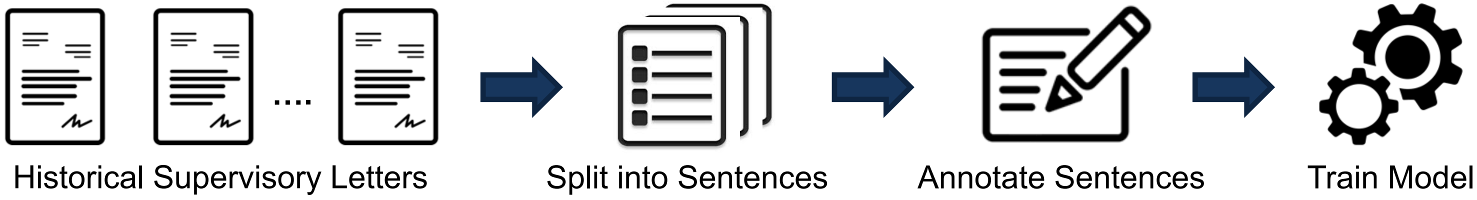
- Privileged access dashboard for easy retrieval of past letters and holistic visualization of letter tones across time and across FIs.
- Able to monitor and spot anomalies in letter tones across FIs of similar risk ratings or of a particular FI across time.



Time series of document-level sentiment scores

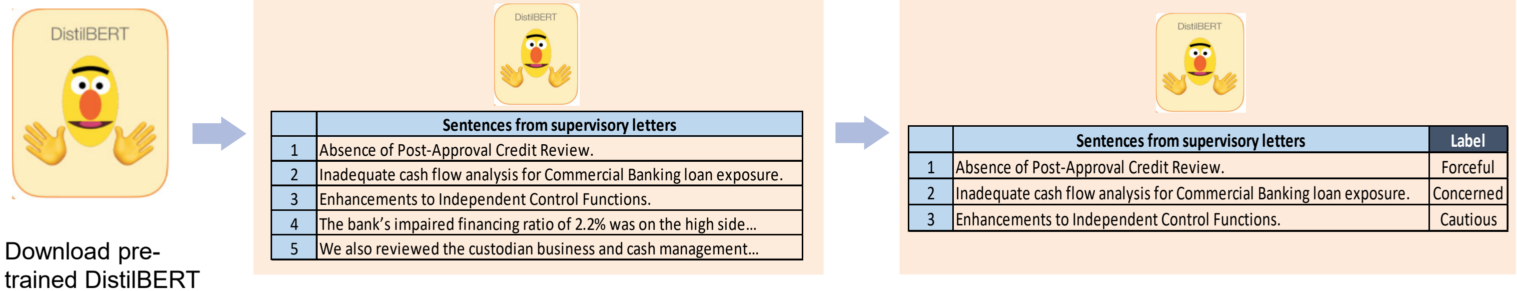
Tone Analysis (1): Methodology

- Multi-class text classification on each sentence in a draft supervisory letter.
- Each sentence is classified into one of 4 ordinal tone classes: ‘Neutral’, ‘Cautious’, ‘Concerned’, ‘Forceful’.
- 15K sentences, labelled by 12 independent supervisors in 4 groups of 3 with quality checks by 3 supervisors on a sampling basis.
- Each sentence contributes a score corresponding to their predicted tone that is aggregated and normalized to arrive at a compound score representing the overall tone of the letter.
- Letters are then assigned to different categories based on overall tone scores for document-level comparisons.



Tone Analysis (2): Model Training and Performance

- Leverage transfer learning to improve model performance.



	Sentences from supervisory letters
1	Absence of Post-Approval Credit Review.
2	Inadequate cash flow analysis for Commercial Banking loan exposure.
3	Enhancements to Independent Control Functions.
4	The bank's impaired financing ratio of 2.2% was on the high side…
5	We also reviewed the custodian business and cash management…

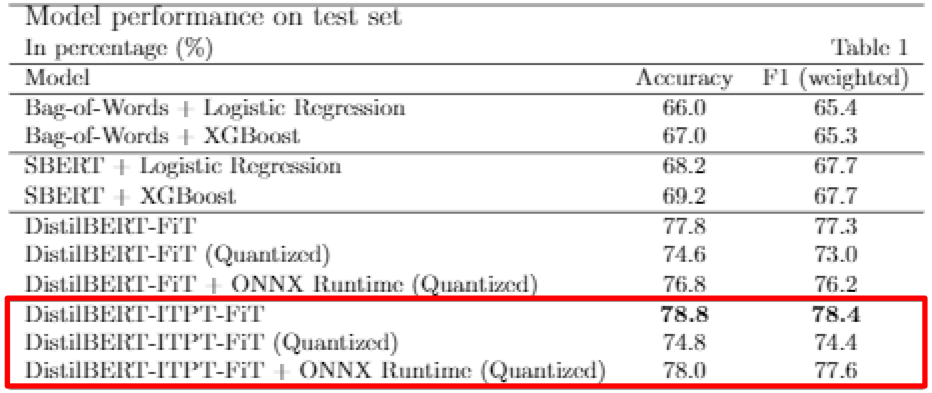
	Sentences from supervisory letters	Label
1	Absence of Post-Approval Credit Review.	Forceful
2	Inadequate cash flow analysis for Commercial Banking loan exposure.	Concerned
3	Enhancements to Independent Control Functions.	Cautious

Download pre-trained DistilBERT from HuggingFace.

Further train the model with all unlabelled sentences.

Fine-tune the model on labelled sentences.

- The fined-tuned transformer-based model performs the best with a weighted-F1 score of 78.4.



Model performance on test set

In percentage (%) Table 1

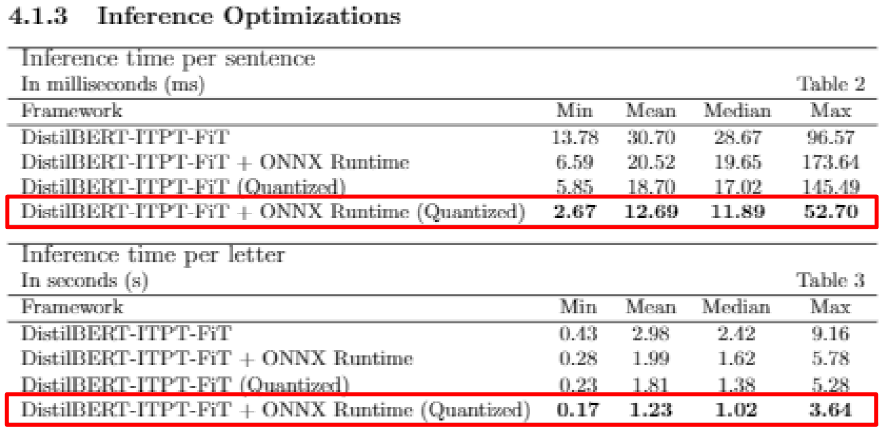
Model	Accuracy	F1 (weighted)
Bag-of-Words + Logistic Regression	66.0	65.4
Bag-of-Words + XGBoost	67.0	65.3
SBERT + Logistic Regression	68.2	67.7
SBERT + XGBoost	69.2	67.7
DistilBERT-FiT	77.8	77.3
DistilBERT-FiT (Quantized)	74.6	73.0
DistilBERT-FiT + ONNX Runtime (Quantized)	76.8	76.2
DistilBERT-ITPT-FiT	**78.8**	**78.4**
DistilBERT-ITPT-FiT (Quantized)	74.8	74.4
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	78.0	77.6

Accuracy and F1 score on test data

Image source: https://jalammar.github.io/illustrated-transformer/

Tone Analysis (3): Inference Optimizations

- Utilize ONNX Runtime and quantization to optimize inference speed.
- Reduced median inference time for a sentence by 2.4 times from 28.67 milliseconds to **11.89 milliseconds**.
- Reduced maximum inference time for a letter by 2.5 times from 9.16 seconds to **3.64 seconds**.
- Minimal impact to the weighted-F1 score (decreased by 0.8%).



4.1.3 Inference Optimizations

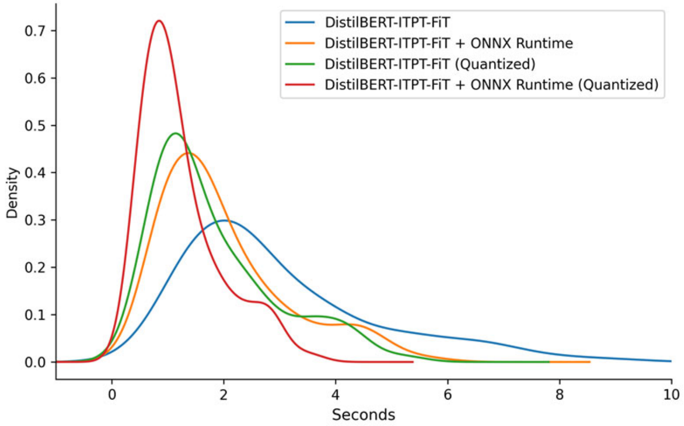
Inference time per sentence
In milliseconds (ms) — Table 2

Framework	Min	Mean	Median	Max
DistilBERT-ITPT-FiT	13.78	30.70	28.67	96.57
DistilBERT-ITPT-FiT + ONNX Runtime	6.59	20.52	19.65	173.64
DistilBERT-ITPT-FiT (Quantized)	5.85	18.70	17.02	145.49
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	**2.67**	**12.69**	**11.89**	**52.70**

Inference time per letter
In seconds (s) — Table 3

Framework	Min	Mean	Median	Max
DistilBERT-ITPT-FiT	0.43	2.98	2.42	9.16
DistilBERT-ITPT-FiT + ONNX Runtime	0.28	1.99	1.62	5.78
DistilBERT-ITPT-FiT (Quantized)	0.23	1.81	1.38	5.28
DistilBERT-ITPT-FiT + ONNX Runtime (Quantized)	**0.17**	**1.23**	**1.02**	**3.64**

Model inference time

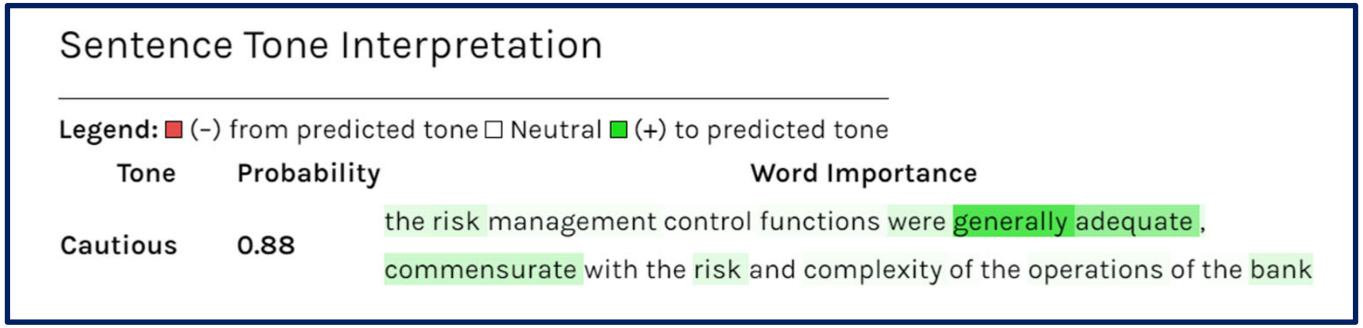


Inference time distribution for a letter

BANK NEGARA MALAYSIA
CENTRAL BANK OF MALAYSIA

Tone Analysis (4): Model Prediction Interpretation

- Utilize integrated gradients in Captum to identify words (in green) that significantly contribute towards the predicted sentence tone. Higher color intensity indicates higher contribution and vice versa.
- Display probability as a confidence score for the model prediction.

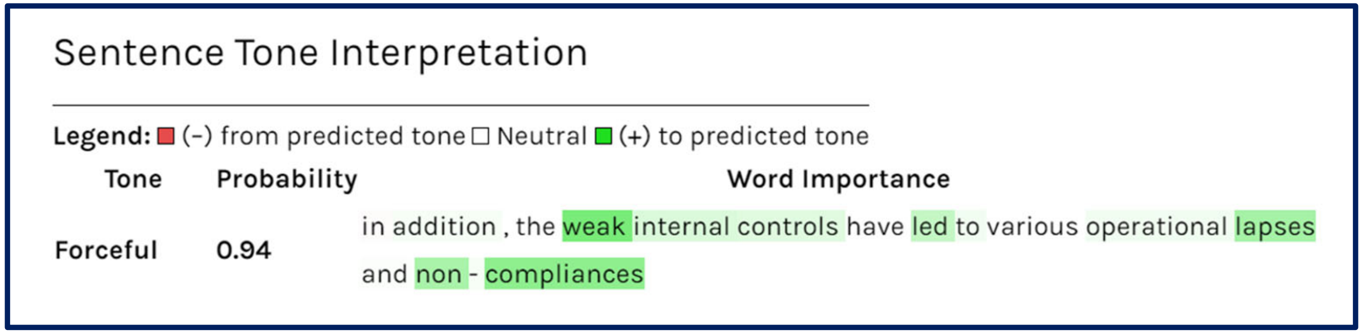


Sentence Tone Interpretation

Legend: 🟥 (-) from predicted tone □ Neutral 🟩 (+) to predicted tone

Tone	Probability	Word Importance
Cautious	0.88	the risk management control functions were generally adequate , commensurate with the risk and complexity of the operations of the bank

Interpretation for a sentence with predicted "cautious" tone



Sentence Tone Interpretation

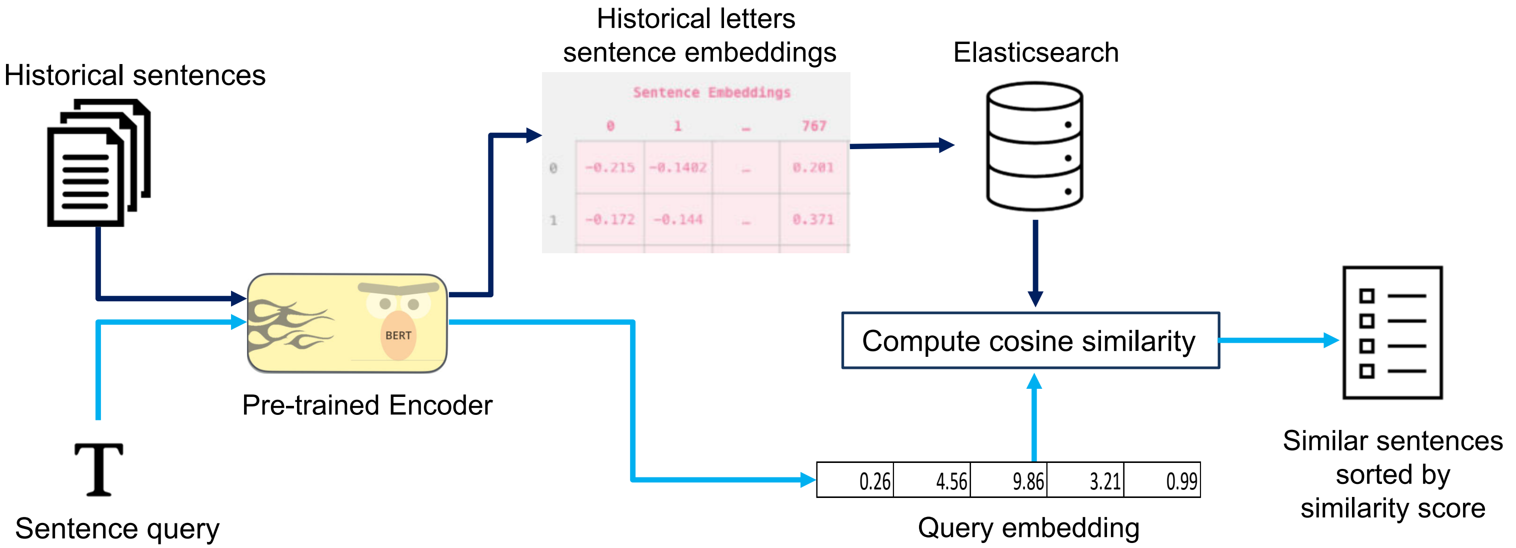
Legend: 🟥 (-) from predicted tone □ Neutral 🟩 (+) to predicted tone

Tone	Probability	Word Importance
Forceful	0.94	in addition , the weak internal controls have led to various operational lapses and non - compliances

Interpretation for a sentence with predicted "forceful" tone

Sentence Search: Approach

- Utilize a pre-trained Sentence-BERT model ("all-mpnet-base-v2").
- Encode sentences from all historical supervisory letters into sentence embeddings, then store in Elasticsearch.
- Upon receiving a sentence query, the model encodes the raw text into an embedding then calculates the cosine similarities between the query embedding and the sentence embeddings from historical supervisory letters.
- The results are then presented in descending order by similarity scores.


Historical letters sentence embeddings
Elasticsearch
Historical sentences
BERT
Pre-trained Encoder
Compute cosine similarity
T
Sentence query
Query embedding
Similar sentences sorted by similarity score

Sentence Embeddings

	0	1	…	767
0	-0.215	-0.1402	…	0.201
1	-0.172	-0.144	…	0.371

0.26	4.56	9.86	3.21	0.99

Image source: https://jalammar.github.io/illustrated-transformer/