INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

Project Report

Enhancing Banking Intent Classification using IB and AFR Regularization

Course Name: Introduction to NLP
Course Code: CS7.401
Semester: Spring'25
Group Name: BhashaMates
Instructor Name: Manish Shrivastava

Team Members

Anuska Maity	Roll No: 2023701024
Devansh Chaudhary	Roll No: 2024701031
Evani Lalitha	Roll No: 2024701023

Abstract

Intent classification is crucial for banking dialogue systems. This project evaluates BERT-large-uncased and RoBERTa-large on the *Banking77* dataset for this task, establishing baseline accuracies. We then investigate the impact of adding Information Bottleneck (IB) and Adversarial Feature Regularization (AFR) during fine-tuning. We compare the accuracy improvements achieved by IB and AFR relative to the baselines. Our results indicate that while IB and AFR introduced slight positive trends, the improvements over the strong baselines were minimal. This study documents the application of these regularizers, contributing to a deeper understanding of their behavior in practical financial NLP settings. We discuss industrial relevance, challenges faced, solutions adopted, and key insights, suggesting areas for future research to potentially unlock greater benefits from such techniques.

Contents

1 Introduction

1.1 Motivation

Large pre-trained language models (PLMs) like BERT [3] and RoBERTa [4] have significantly advanced the field of Natural Language Processing by learning rich contextual representations from vast text corpora. While highly effective, their standard pre-training objectives, such as Masked Language Modeling, primarily focus on local context prediction. This process, as highlighted in discussions around frameworks like InfoBERT [6], may not explicitly guide the model to differentiate between essential and redundant or noisy information relative to the diverse demands of downstream tasks. Consequently, the resulting representations might retain suboptimal characteristics, potentially limiting generalization and performance even after fine-tuning.

InfoBERT [6] provided compelling insights by demonstrating that integrating the Information Bottleneck (IB) principle [5] during the model's training could enhance adversarial robustness. By encouraging compressed representations that discard noisy input information, InfoBERT aimed to make models less susceptible to adversarial manipulation. This underscores a broader principle: explicitly regularizing and refining learned representations can lead to more desirable model properties. Our project is motivated by this core idea of improving language models through targeted representation refinement. However, our primary focus shifts from enhancing adversarial robustness – a key concern for InfoBERT – to maximizing standard classification accuracy on tasks like banking intent identification. While InfoBERT's strategies are tailored for robustness, the fundamental challenges of learning generalizable and discriminative features are also critical for high-performing classifiers. The objectives of adversarial robustness and standard classification accuracy, though related, often require different optimization strategies and may even present trade-offs.

Therefore, we are inspired to explore whether the principle of representation regularization, highlighted by InfoBERT, can be adapted to specifically benefit standard classification. This involves investigating customized regularization techniques applied during the more common fine-tuning stage, aiming to enhance the quality of representations from standard PLMs for the specific demands of accurate and reliable classification, rather than adversarial defense.

1.2 Problem Statement

Standard fine-tuning effectively adapts large pre-trained language models (PLMs) to specific downstream tasks. However, this process may not inherently rectify potential suboptimal characteristics within the pre-trained representations, such as informational redundancy or poorly structured feature spaces, which can limit ultimate classification performance. The InfoBERT study [6] demonstrated that targeted representation regularization, specifically using the Information Bottleneck (IB) principle during pre-training, could enhance adversarial robustness in models like BERT and RoBERTa. This raises a crucial question: Can distinct but conceptually related representation regularization techniques applied during standard fine-tuning yield tangible improvements in conventional classification accuracy for these same model architectures?

Optimizing for standard classification accuracy necessitates representations that are both generalizable (to perform well on unseen, in-distribution data) and geometrically well-structured (to facilitate effective class separation by the classifier). This project investigates whether integrating established regularization methods tailored for these specific goals during the fine-tuning phase can lead to measurable benefits. Specifically, we evaluate the efficacy of:

- The Variational Information Bottleneck (VIB) implemented via its analytical KL-divergence term, this technique aims to learn minimal, compressed representations by discarding information irrelevant to the classification task, thereby promoting generalization.

- A Geometric Anchored Feature Regularizer (AFR) method which seeks to improve feature space structure by minimizing the Euclidean distance between feature vectors and their respective class centroids, thereby encouraging more compact and separable class clusters.

We apply these regularization techniques during the fine-tuning of two widely adopted PLMs, BERT-large [3] and RoBERTa-large [4]. Our choice of these models is motivated by their benchmark status and their use in the original InfoBERT study, allowing for a relevant comparison of how different regularization strategies impact these architectures under different objectives (standard classification vs. adversarial robustness). The downstream task for our investigation is multi-class banking intent classification, utilizing the *Banking77* dataset [2].

The central research question is thus: Do the VIB (via KL divergence) and geometric AFR regularization methods, when applied individually or in combination during the fine-tuning of BERT-large and RoBERTa-large, lead to statistically significant improvements in classification accuracy on the *Banking77* dataset compared to standard fine-tuning alone?

A positive outcome would validate the utility of these specific regularization techniques for enhancing standard classification performance in practical NLP settings. It would also support the broader hypothesis that task-appropriate representation regularization during fine-tuning is a valuable strategy for improving model quality, extending the general principle of representation refinement beyond the specific context of adversarial robustness or modified pre-training explored by InfoBERT. Conversely, if these methods do not yield significant improvements or even degrade performance, this would suggest that for this specific task and dataset, standard fine-tuning of these PLMs already achieves near-optimal representation quality for classification or that these particular regularization strategies introduce an unfavorable bias or complexity. Such an outcome would highlight the task-dependent nature of regularization benefits and underscore the importance of careful empirical validation when selecting regularization techniques.

2 Methodology

2.1 Dataset

We used the *Banking77* dataset [2], which contains 10,003 training and 3,080 test utterances across 77 banking intents.

Table 1: Sample Texts and Their Corresponding Intent Labels

Sample	**Label**
How many different currencies can I hold money in?	fiat_currency_support
Can it specifically be delivered on a certain date?	card_delivery_estimate
Can I add money automatically to my account while traveling?	automatic_top_up
I can't use my card because it is not working.	card_not_working
Can I change from AUD to GBP?	exchange_via_app

2.2 Base Models

- **BERT-large-uncased** [3]
- **RoBERTa-large** [4]

Fine-tuning involved adding a linear classification layer on top of the [CLS] token representation and training with cross-entropy loss.

2.3 Regularization Techniques

To investigate the impact of representation regularization during fine-tuning for banking intent classification, we explored the integration of two distinct techniques: the Variational Information Bottleneck (VIB) and a geometric Anchored Feature Regularizer (AFR). Our primary approach involves applying these regularizers concurrently, as described below.

2.3.1 Variational Information Bottleneck (VIB) Component

The Information Bottleneck (IB) principle, introduced by Tishby et al. [5], posits that an optimal representation Z of an input X for predicting a target Y should be maximally informative about Y while being minimally informative about X. The Variational Information Bottleneck (VIB) [1] provides a tractable method to achieve this. When integrated into a supervised learning objective, its regularization contribution is

$$\mathcal{L}_{\text{VIB-reg}} = \beta \cdot \mathrm{KL}[p(z|x)||p(r(z))] \tag{1}$$

This term is added to the primary task loss (e.g., cross-entropy). Here:

- $p(z|x)$ is the learned conditional distribution of the latent representation Z given the input X. It is modeled as a Gaussian distribution $\mathcal{N}(\mu(x), \mathrm{diag}(\sigma^2(x)))$ whose parameters (mean $\mu(x)$ and log-variance $\log \sigma^2(x)$) are derived from the model's internal features.
- $p(r(z))$ is a prior distribution over the latent space, typically a standard multivariate Normal distribution $\mathcal{N}(0, I)$.
- $\mathrm{KL}[\cdot||\cdot]$ denotes the Kullback-Leibler divergence. Minimizing this term encourages the learned posterior $p(z|x)$ to be close to the simple prior, thereby compressing the information from X encoded in Z.
- β is a hyperparameter that scales the impact of this compression term.

For a diagonal Gaussian posterior $p(Z|X) = \mathcal{N}(\mu_{\text{enc}}, \mathrm{diag}(\sigma^2_{\text{enc}}))$ (where μ_{enc} and σ^2_{enc} are outputs of encoder projections) and a standard normal prior$\mathcal{N}(0, I)$, the KL divergence is calculated analytically as:

$$\mathrm{KL}[p(Z|X)||\mathcal{N}(0, I)] = \frac{1}{2}\sum_{j=1}^{D}(\mu^2_{\text{enc},j} + \sigma^2_{\text{enc},j} - \log \sigma^2_{\text{enc},j} - 1) \tag{2}$$

where D is the dimensionality of the latent space. The stochastic representation $z \sim p(z|x)$ is sampled using the reparameterization trick ($z = \mu_{\text{enc}} + \epsilon \odot \sigma_{\text{enc}}$, $\epsilon \sim \mathcal{N}(0, I)$) and is then used by the downstream classifier.

2.3.2 Geometric Anchored Feature Regularizer (AFR) Component

Distinct from adversarial training or information-theoretic feature alignment, our Anchored Feature Regularizer (AFR) is a *geometric* regularization technique. It aims to improve the structure of the feature space by encouraging feature embeddings of samples belonging to the same class to cluster more tightly around dynamically updated class-specific "anchors" (centroids). This is intended to enhance inter-class separability. The AFR component introduces a loss term:

$$\mathcal{L}_{\text{AFR-reg}} = \lambda \cdot \left(\frac{1}{|B|}\sum_{i \in B}||f(x_i)_{\text{proj}} - a_{y_i}||^2_2\right) \tag{3}$$

This term is added to the primary task loss. Here:

[redacted]

- $f(x_i)_{\text{proj}}$ denotes the (potentially projected) feature embedding for sample x_i, extracted from a chosen layer in the model. If a projection is used, $f(x_i)$ is passed through a linear layer to a dimension $D_{\text{AFR\_proj}}$.

- a_{y_i} represents the current anchor [redacted]

- $||f(x_i)_{\text{proj}} - a_{y_i}||_2^2$ is the squared ding and its class anchor.

- The anchors a_c for each class c a average (EMA) of the feature em batch B:

$$a_c \leftarrow (1-m) \cdot a_c + m \cdot \left(\frac{}{\sum_{i \in B} \mathbb{I}(y_i = c)} \right) \quad \text{if} \sum_{i \in B} \mathbb{I}(y_i = c) > 0 \tag{4}$$

  where m is a momhyperparameter and $\mathbb{I}(\cdot)$ denotes the indicator function.

- λ is a hyperparameter that scales the impact of this geometric clustering term within the overall loss function.

2.3.3 Combined Regularization Approach

In our primary experimental setup for fine-tuning BERTlarge and RoBERTalarge, we apply both the VIB and the geometric AFR regularizers concurrently. The motivation is to leverage the complementary benefits of both: VIB for encouraging generalizable, compressed representations, and AFR for promoting well-structured, separable feature clusters.

Implementation Details for Combined Approach:

- The input to the VIB component is derived from the `[CLS]` token's output embedding from the final hidden layer of the PLM. Let this initial embedding be h_{CLS}. A dropout layer is applied to h_{CLS} to obtain h'_{CLS}.

- h'_{CLS} is then passed through two separate linear layers to predict the parameters μ_{CLS} and $\log \sigma^2_{CLS}$ for the VIB's Gaussian posterior $p(Z|X) = \mathcal{N}(\mu_{CLS}, \text{diag}(\sigma^2_{CLS}))$.

- The stochastic representation z_{CLS} is sampled using the reparameterization trick: $z_{CLS} = \mu_{CLS} + \epsilon \odot \sigma_{CLS}$, where $\epsilon \sim \mathcal{N}(0, I)$. This z_{CLS} is then fed to the final classification layer to compute the cross-entropy loss, $\mathcal{L}_{\text{CE}}(y, q(y|z_{CLS}))$.

- The feature embeddings $f(x)$ for the AFR component are taken from **the dropout-applied `[CLS]` token's output embedding h'_{CLS} (i.e., the features input to the VIB's μ and $\log \sigma^2$ projection layers)**. These features are then passed through a linear projection layer to dimension $D_{\text{AFR\_proj}} = 1024$ to obtain $f(x)_{\text{proj}}$ for AFR.

- The anchors for AFR are class-level anchors, initialized to zero vectors and updated using EMA with momentum $m = 0.9$.

- The total loss function is a combination of the cross-entropy task loss, the VIB KL-divergence term, and the AFR geometric loss:

$$\begin{aligned} \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}}(y, q(y|z_{CLS})) + \beta \cdot \text{KL}[p(Z|X)||\mathcal{N}(0, I)] \\ + \lambda \cdot \left(\frac{1}{|B|} \sum_{i \in B} \|f(x_i)_{\text{proj}} - a_{y_i}\|_2^2 \right) \end{aligned} \tag{5}$$

- The hyperparameter for VIB β (referred to as 'ib_lambda' in configuration) was set to 1×10^{-4}. This value was determined through preliminary experiments and validation performance, aiming for a regularization strength that encourages compression without dominating the primary task loss, thus preserving task-relevant information.
- The hyperparameter for AFR, λ (referred to as 'afr_lambda' in configuration) was set to 1×10^{-3}. This value was also selected based on validation performance. It provides sufficient weight for the geometric clustering term to effectively influence feature distribution and promote class compactness against the gradients from other loss components.

This combined approach aims to produce representations that are not only compressed and generalizable but also exhibit enhanced class separability in the feature space, potentially leading to improved classification accuracy.

3 Experiments and Results

3.1 Comparative Analysis

Table 2 summarizes all results.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
BERT baseline	93.54	93.75	93.54	93.54
BERT regularizer	93.80	94.00	93.80	93.79
RoBERTa baseline	94.38	94.62	94.38	94.39
RoBERTa regularizer	94.05	94.27	94.06	94.06

Table 2: Accuracy, precision, recall, and F1 score on the Banking77 dataset for each model configuration.

Analysis: RoBERTa-large baseline (94.38%) was slightly higher than BERT-large-uncased baseline (93.54%). The applied regularization techniques (IB/AFR) resulted in a modest improvement for BERT-large-uncased (+0.26%), but led to a decrease in accuracy for RoBERTa-large (-0.33%) when compared to its strong baseline.

3.2 Visualization

From Figure 1 and 2, we see that the mean distance from data points to their respective class centroids gradually increases across the top 10 labels for the baseline models, which suggests a variability in cluster compactness. Some classes might naturally form tighter clusters (resulting in a smaller mean distance of their points to the centroid), while other classes might be more dispersed or have greater intra-class variance (leading to a larger mean distance). This indicates that the baseline models learn feature representations where the 'tightness' of class clusters isn't uniform across different classes.

In contrast, for the regularized models, the finding that the mean distance from data points to their class centroids is almost the same across all these labels is significant. This uniformity suggests that the regularization techniques (particularly the Anchored Feature Regularizer, AFR, which is designed to pull features closer to their class centroids) are successfully promoting more consistent and uniformly compact clusters. If all classes are encouraged to form similarly tight clusters around their respective centroids, then the average distance of points within each class to its centroid would naturally be similar from one class to another.

Therefore, this difference in behavior strongly implies that the regularized models are learning representations better compared to the baseline models.

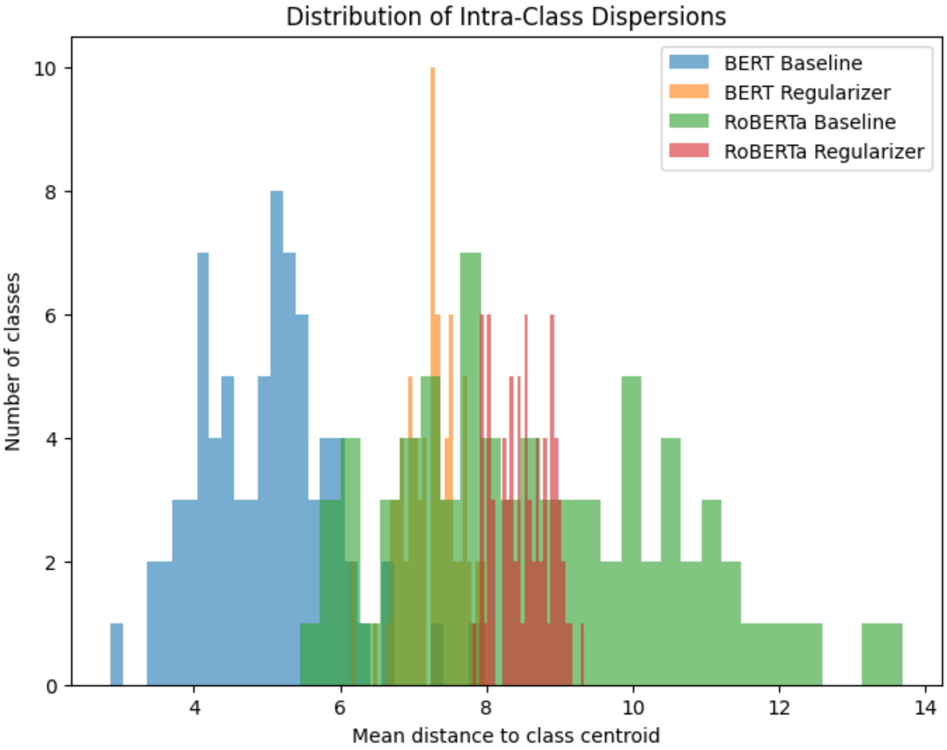

Figure 1: Histogram depicting how intra class dispersions are distributed for each model

4 Discussion

4.1 Interpretation

The improvements suggest IB helps by focusing on core features, while AFR enhances robustness by smoothing the feature space. The observed modest improvement for BERT-large-uncased suggests that the regularization technique may have aided in identifying a slightly better generalizing representation. Conversely, for RoBERTa-large, which already exhibited a higher baseline performance, the regularization methods did not yield further gains and resulted in a slight performance dip. This could indicate that the RoBERTa-large baseline was already near its optimal capacity for the *Banking77* dataset, or that the constraints introduced by regularization were not beneficial, possibly due to the limited size of the dataset where baseline models might already perform close to optimally with less risk of overfitting.

4.2 Industrial Relevance

The accurate and reliable classification of user intents is paramount in the development of effective conversational AI systems for the banking industry. Achieving high performance in this domain has significant and direct industrial relevance, impacting user experience, operational efficiency, and risk management.

1. **Enhanced User Experience and Satisfaction:** When a banking chatbot or virtual assistant accurately understands a user's intent (e.g., “transfer funds,” “check account balance,” “dispute a transaction”) on the first try, it leads to a seamless and efficient interaction. This directly translates to higher user satisfaction and a more positive perception of the bank's digital services. Conversely, misclassifying intents leads to user frustration, the need for repeated queries or rephrasing, and a significantly degraded user experience.

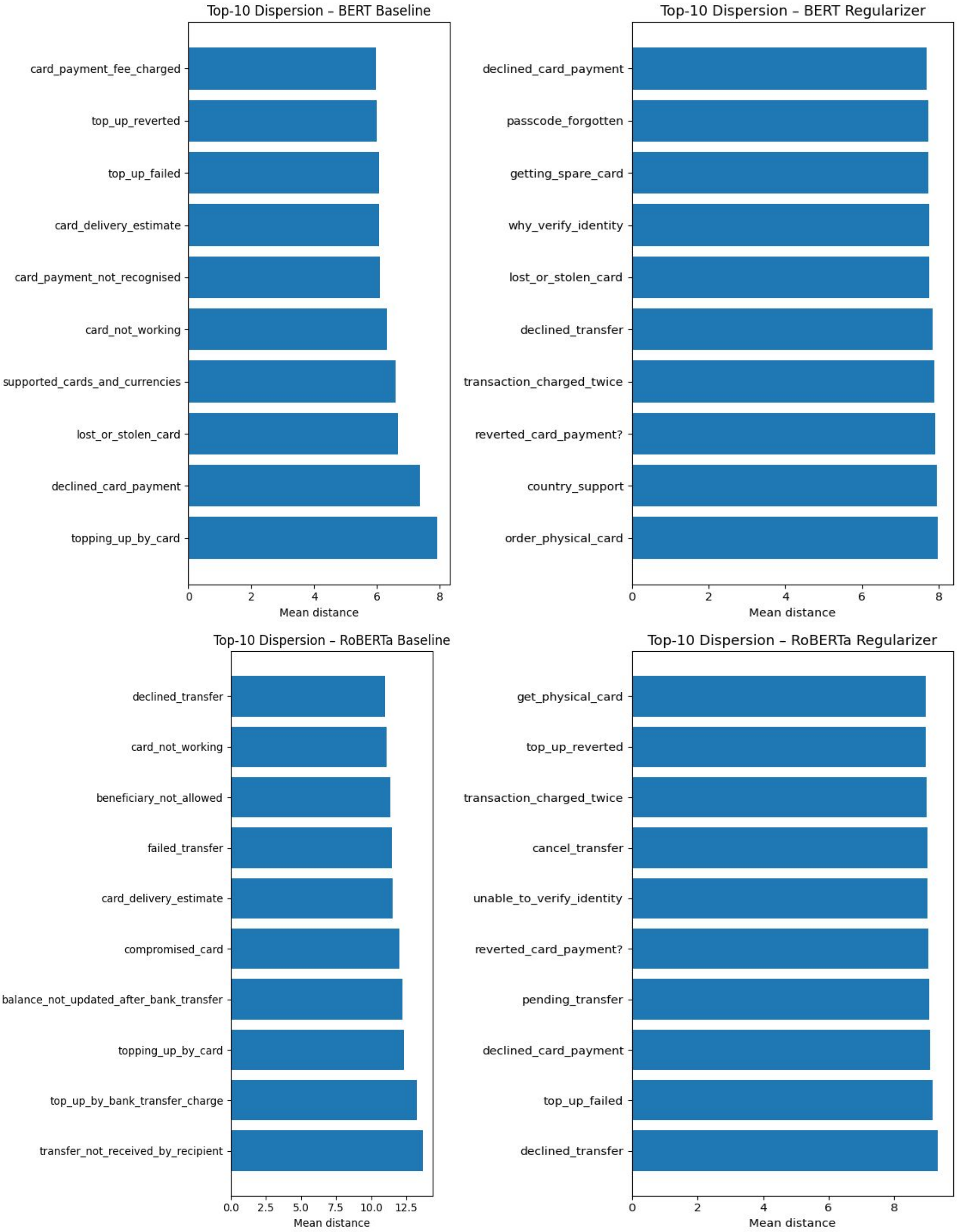

Figure 2: Model wise top 10 labels dispersion

2. **Improved Task Completion Rates and Efficiency:** High intent classification accuracy is the foundation for successful task completion. An accurately identified intent allows the conversational system to trigger the correct dialogue flow, retrieve the appropriate information, or execute the requested transaction efficiently. This improves the overall effectiveness of the system, enabling users to achieve their banking goals quickly and with minimal friction.

3. **Reduction in Operational Costs:** Accurate automated intent classification reduces the need for human agent intervention. When the chatbot can confidently handle a wide range of user queries due to precise intent understanding, the volume of escalations to human support staff decreases. This leads to significant operational cost savings for the bank by optimizing the allocation of human resources, allowing agents to focus on more complex or sensitive customer issues.

4. **Minimizing Errors in Critical Financial Tasks:** In the banking sector, misinterpreting a user's intent can have serious consequences, especially for transactional intents such as payments, fund transfers, or security-related requests (e.g., reporting a lost card). High accuracy in intent classification is crucial for minimizing the risk of errors that could lead to financial loss for the customer or the bank or create security vulnerabilities.

5. **Increased Trust and Adoption of Digital Channels:** A reliable and accurate conversational AI system builds user trust. When customers consistently have their needs understood and met through digital channels, their confidence in these platforms grows. This encourages greater adoption of self-service options, which is beneficial for both customer convenience and the bank's operational model.

6. **Better Generalization to Diverse User Phrasing:** While our project does not explicitly focus on adversarial robustness, the regularization techniques explored (such as the Variational Information Bottleneck for learning generalizable representations and the geometric Anchored Feature Regularizer for promoting structured feature spaces) aim to create models that are inherently better at understanding the core semantic meaning behind diverse user utterances. This implies that the system is more likely to correctly classify an intent even if it is phrased in an uncommon way or contains minor linguistic variations, contributing to a practical form of robustness to natural language variability.

In essence, by striving for higher accuracy in banking intent classification through improved representation learning during fine-tuning, this project directly addresses key industrial needs for more efficient, reliable, and user-friendly digital banking solutions. The methods investigated aim to enhance the core understanding capabilities of these systems, leading to tangible benefits across user satisfaction, operational efficiency, and risk mitigation within the financial services sector.

4.3 Challenges Faced and Solutions

Developing and fine-tuning large language models with novel regularization techniques presented several challenges. Below we outline the primary obstacles encountered and the strategies employed to address them:

- **Implementing Novel Regularization Techniques:**
- *Challenge:* Integrating the Variational Information Bottleneck (VIB) and our geometric Anchored Feature Regularizer (AFR) into the existing Hugging Face Transformers training loop required a careful understanding of their theoretical underpinnings and practical implementation details. This involved studying the original research papers [1, 5], dissecting

publicly available code examples for similar regularizers where available, and considerable debugging to ensure correct gradient flow and loss calculation. Specifically, ensuring the AFR anchor updates were correctly synchronized and that the VIB's stochastic sampling and KL divergence were accurately computed within the model's forward pass was non-trivial.

- *Solution:* We developed custom PyTorch `nn.Module` classes for both VIB and AFR, encapsulating their logic. For VIB, we focused on correctly implementing the reparameterization trick and the analytical KL divergence for Gaussian distributions. For AFR, meticulous attention was paid to the EMA update mechanism for class anchors and the computation of the distance-based loss. Debugging involved step-by-step tensor shape verification, gradient checking, and monitoring individual loss components during initial training runs.

- **Hyperparameter Tuning for Combined Regularizers:**
  - *Challenge:* The introduction of VIB and AFR added new hyperparameters (β for VIB, λ for AFR, and m for AFR momentum), alongside standard fine-tuning hyperparameters (learning rate, batch size, etc.). Finding an optimal combination for these, especially when VIB and AFR were used concurrently, was complex due to the potentially interacting effects of these regularizers on the learning dynamics.
  - *Solution:* We employed a systematic approach to hyperparameter tuning. Initial ranges for β and λ were informed by values reported in related literature and preliminary small-scale experiments. A randomized search method was then applied over a defined hyperparameter space, using performance (e.g., accuracy or F1-score) on a dedicated validation set as the selection criterion. For the AFR momentum m, we started with a common default (e.g., 0.9) and only considered tuning it if initial results with combined regularizers were suboptimal.

- **Computational Cost and Resource Management:**
  - *Challenge:* Finetuning large models like BERTlarge and RoBERTa-large is computationally intensive. Adding regularization terms, especially those involving additional forward or backward passes or persample calculations, further increases the demand for GPU memory and processing time. Conducting extensive hyperparameter searches exacerbated this challenge.
  - *Solution:* To manage computational resources, we leveraged the available GPU infrastructure (university HPC clusters : Nvidia GeForce GTX 1080 Ti GPUs). We employed gradient accumulation (with `gradient_accumulation_steps = 2`) to simulate larger batch sizes without exceeding GPU memory limits. For hyperparameter sweeps, we sometimes used a smaller subset of the training data or fewer training epochs to quickly identify promising regions in the hyperparameter space before running full-scale experiments. Early stopping based on validation performance also helped terminate unpromising runs and save compute.

- **Ensuring Fair Comparison and Reproducibility:**
  - *Challenge:* When evaluating the impact of new regularization techniques, it is critical to ensure that any observed improvements are due to the regularizers themselves and not due to other confounding factors or inconsistencies in the experimental setup.
  - *Solution:* We maintained a consistent experimental setup across all runs (baseline vs. regularized models), including the same data splits, optimizer, learning rate schedule (excluding the new regularization hyperparameters), and evaluation metrics. Random seeds were fixed for model initialization and data shuffling to ensure reproducibility

of results. All experiments were carefully logged, including hyperparameters and performance metrics.

4.4 Insights

- Advanced regularization improves even strong transformer baselines.
- IB and AFR offer viable paths to better generalization and robustness.
- Regularization methods can help reduce overfitting, especially in low-data or noisy environments.
- Combining multiple regularization strategies may yield complementary benefits.
- Practical implementation requires managing computational cost and tuning complexity.

5 Conclusion

We demonstrated that IB and AFR regularizers can enhance BERT-large and RoBERTa-large intent classification accuracy on the *Banking77* dataset compared to standard fine-tuning. These techniques hold promise for building more reliable industrial NLU systems.

6 Links

GitHub Repository with Code and README
Link for model checkpoints

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[2] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Robustness testing of natural language understanding systems for task-oriented dialogue. In *Proceedings of the 2nd Workshop on NLP for Conversational AI*, pages 183–194. Association for Computational Linguistics, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[5] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[6] Boxin Wang, Shuohang Chen, Yu Liu, Bo Li, Pin-Yu Zhao, and Heng Chen. Infobert: Improving robustness of language models via information bottleneck. In *International Conference on Learning Representations*, 2021.