**FACULTY OF COMPUTER SCIENCE AND ENGINEERING**

**National University of Computer and Emerging Science, Islamabad**

**Subject:** [**Adv. ML / ML for DS**](https://classroom.google.com/u/0/c/NzQ2MTMxOTU0Njk2) **(MS-DS) Instructor: Dr. M. Ishtiaq**

## **Project Proposal**

## 1. Project Title

Understanding Emotions More Clearly: Elevating Data Quality for Better Emotion Classification.

## 2. Team Information

|  |  |
| --- | --- |
| Student Name: Muhammad Usman Arif  Registration Number: 24i-8001  Email [i248001@isb.nu.edu.pk](mailto:i248001@isb.nu.edu.pk) | Student Name: Syed Shahbaz Ali Shah  Registration Number 24i-8002  Email [i248002@isb.nu.edu.pk](mailto:i248002@isb.nu.edu.pk) |
| Student Name: Ammar Ahmed Siddique  Registration Number(s): 24i-8057  Email: [i248057@isb.nu.edu.pk](mailto:i24@isb.nu.edu.pk) | Student Name: Muhammad Shahzad  Registration Number: 24i-8053  Email: [i248053@isb.nu.edu.pk](mailto:i248053@isb.nu.edu.pk) |

## 3. Problem Statement

| **Aspect** | **Details** |
| --- | --- |
| **Context of the Problem** | This research proposal focuses on the fine-grained classification of emotions in natural language using the GoEmotions dataset (Demszky et al., 2020). The dataset contains approximately 58,000 Reddit comments, each annotated with one or more of 28 emotion categories and an additional “neutral” label. |
| **Importance** | **Sentiment Analysis**: Enables organizations to understand nuanced emotional expressions in customer feedback across social media and product reviews.  **Mental Health Monitoring**: Assists in detecting subtle emotional cues in user-generated content, particularly for well-being and therapeutic applications.  **Human–Computer Interaction**: Supports the development of empathetic systems, such as emotionally intelligent chatbots and adaptive tutoring platforms. |
| **Challenges** | **Class Imbalance**: Dominant categories such as “neutral” (~11,000 samples) vastly outnumber underrepresented emotions like “pride” (~300 samples), often causing classifiers to ignore minority classes.  **Semantic Drift in Augmentation**: Synthetic data generation may inadvertently alter the intended emotional tone or introduce noise. |

## 4. Objectives

1. **Mitigate Class Imbalance via Targeted Augmentation**  
   To address class imbalance, we propose the implementation and comparative analysis of augmentation strategies: synonym substitution. These methods aim to enhance the representation of under-represented emotion labels within the GoEmotions dataset.
2. **Quantify Augmentation Impact on Model Performance**  
   To enhance our analysis, we will evaluate a pre-fine-tuned BERT classifier on (a) the original GoEmotions dataset, and (b) augmented variant. Following this, we will compare the per-class precision, recall, and F1-scores, with a particular focus on minority emotions.
3. **Analyze Trade-offs Across Augmentation Methods**  
   Assess how each technique impacts the performance of majority and minority labels, the overall macro-F1 score, and the stability of training, including factors such as convergence speed and the potential for overfitting.
4. **Deliver Practical Recommendations for Emotion-Rich NLP**  
   Based on empirical findings, it is advisable to recommend an augmentation approach, or a combination of methods, that strikes the optimal balance between simplicity, semantic fidelity, and improvements in classification performance. This guidance will contribute to the advancement of future emotion-classification pipelines.

## 5. Relevance to Course Topics

1. **Supervised Learning (Classification with Neural Nets)**  
   We have refined a pre-trained BERT model, which is a deep neural network, to categorize sentences into one of 28 specific emotion labels. This process includes a training loop, utilization of the cross-entropy loss function, and the implementation of backpropagation, all of which effectively demonstrate the mechanics of neural network training.
2. **Performance Metrics (Precision, Recall, F1-Score)**  
   We will calculate per-class precision, recall, and F1 scores to evaluate the impact of augmentation on both rare and common emotions. Additionally, we will use macro-F1 scores to assess overall class balance.
3. **Data Imbalance & Robustness via Data Augmentation**  
   Although they do not qualify as a formal "adversarial attack," our focused synonym substitution is akin to robustness techniques. These strategies aim to reduce the classifier's tendency to overfit to the majority classes. This method fits well within the framework of robustness and regularization, as it effectively synthesises additional examples for minority classes.
4. **Representation Learning / Embeddings (Dimensionality Reduction Analogy)**  
   The contextual embeddings generated by BERT create a high-dimensional feature space from which we derive our class predictions. Our method for augmenting contextual word embeddings utilizes the same foundational transformer representations.
5. **Interpretability (Optional)**  
   By comparing confusion matrices and examining which token substitutions enhance or deteriorate minority-class performance, we gain insights into the model’s sensitivities and decision boundaries.
6. **Bayesian Methods & Uncertainty Estimation (Future Extension)**  
   Although not included in the main proposal, there is an opportunity to enhance this work by incorporating Monte Carlo dropout or Deep Ensembles alongside BERT. These techniques could be instrumental in estimating prediction uncertainty, particularly for emotion classes with limited support.

## 6. Dataset(s)

| **Aspect** | **Details** |
| --- | --- |
| **Source** | GoEmotions dataset (Demszky et al., 2020) – GitHub & Google Cloud |
| **Size & Format** | ~58 000 Reddit sentences; JSON/CSV with fields text and multi-label emotions |
| **Features** | Raw comment text |
| **Labels** | 28 fine-grained emotion categories + “neutral” |
| **Class Distribution** | Highly imbalanced—for example:   * neutral: ~11 000 (19%) * gratitude: ~10 000 (17%) * grief: ~450 (0.8%) * pride: ~280 (0.5%) |

## 7. Proposed Methodology

1. **EDA & Preprocessing**
   1. Conduct an analysis of label frequencies and visualize the distribution of these labels.
2. **Data Augmentation**
   1. **Synonym substitution** (WordNet): replace 1–2 words per sentence.
   2. Record multi-label ground-truth metrics (e.g. label-ranking average precision, micro-F1) on the original dataset **before** collapsing to single labels, to quantify information loss.
3. **Dataset Construction**
   1. Split the filtered dataset into train/dev/test (80/10/10) following Demszky et al. (2020).
4. **Model Training**
   1. Evaluate the pre fine-tuned **BERT** model on both the original and augmented datasets.
   2. Configure the evaluation pipeline with a batch-size of 16, ensuring consistent metrics such as precision, recall, and F1-scores for comparison. Perform the analysis under identical conditions to assess the impact of dataset augmentation.
   3. Record and analyze per-class performance, with a particular focus on minority emotions.
   4. **Metrics:** per-class precision, recall, F1.
5. **Evaluation**
   1. Compare per-class precision, recall, and F1, including **macro-F1** against the original baseline.
   2. Focus on minority classes (e.g., grief ↑ 0.71 F1).
   3. Analyze confusion matrices & loss curves.

## 8. Expected Outcomes

| **Outcome** | **Success Indicator** |
| --- | --- |
| **Boost Minority-Class F1** | Increase F1 for rare emotions (e.g., “grief”) from ~0.45 to ≥0.70 |
| **Improve Macro-F1** | Raise macro-averaged F1 by ≥8 points (from ~0.46 to ≥0.54) |
| **Validate Augmentation Methods** | BERT-CWE augmentation yields higher per-class F1 and faster convergence vs. synonym |
| **Actionable Guidelines** | Clear recommendation of augmentation type & volume for balanced datasets |
| **Robustness Insights** | Reduced false negatives on challenging labels (e.g., “pride,” “remorse”) |

## 9. Project Modules

| **Module** | **Deliverables & Tasks** |
| --- | --- |
| **Literature Review & EDA** | Survey Data Augmentation and Emotion Classification Literature  Download and Profile the GoEmotions Dataset (including class counts and text lengths)  Visualize Distribution Imbalances and Token Distributions |
| **Baseline Model** | Evaluate the pre-fine-tuned BERT model using the original dataset. Document baseline performance metrics, including per-class precision, recall, and F1 Scores, as well as the macro-average F1 score. Additionally, ensure the preservation of evaluation metrics, and comparison results for both the original and augmented datasets for future reference. |
| **Evaluation & Documentation** | Conduct a comparative analysis of all model variants utilising confusion matrices and F1 score, etc., analysis. Assess the differences in robustness and convergence. Compile the definitive report, accompanying figures, and code notebook. |

## 10. Challenges & Risk Mitigation

| **Challenge** | **Risk** | **Mitigation** |
| --- | --- | --- |
| **Semantic Drift in Augmentation** | Synonym swaps may alter sentiment. | Restrict replacements to high-confidence WordNet synsets or BERT similarity thresholds.  Manually review a random 1% sample. |
| **Overfitting on Synthetic Data** | Model may memorize artifacts of augmented samples. | Mix original and synthetic data (50/50 max).  Use dropout and early stopping based on validation loss. |
| **Compute & Time Constraints** | Evaluating BERT across multiple dataset variants is resource-intensive. | Leverage batch processing and efficient pipelines.  Use smaller “distil” BERT for prototyping. |
| **Residual Class Imbalance** | Some emotions still under-represented post-augmentation. | Dynamically adjust augmentation volume per label.  Apply weighted loss or focal loss to emphasize hard classes. |
| **Noisy or Irrelevant Augmentations** | Low-quality sentences degrade model performance. | Filter augmented sentences by semantic similarity score (>0.8).  Limit max 5 augmentations per original. |
| **Evaluation Bias** | Macro-F1 may hide majority-class drops. | Report both macro and weighted metrics.  Include per-class confusion matrices in the analysis. |

11. References

1. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, “GoEmotions: A dataset of fine-grained emotions,” *arXiv preprint arXiv:2005.00547*, 2020.
2. J. Wei and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks,” *arXiv preprint arXiv:1901.11196*, 2019.
3. S. Kobayashi, “Contextual augmentation: Data augmentation by words with paradigmatic relations,” in *Proceedings of NAACL-HLT 2018*, pp. 452–457.
4. D. Mahata, R. R. Shah, M. Kulkarni, and A. Awadallah, “Improving sentiment classification in noisy data using data augmentation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 5724–5731.
5. M.-L. Zhang and Z.-H. Zhou, “A review on multi-label learning algorithms,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.
6. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, et al., “RoBERTa: A robustly optimized BERT pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
7. R. Saini, P. Gupta, and S. Kumar, “Hierarchical attention networks for fine-grained emotion classification,” in *Proceedings of ACL 2021*, pp. 1148–1158.
8. S. Akhtar, D. Gupta, A. Ekbal, and P. Bhattacharyya, “A multi-task and transfer learning approach for emotion detection and sentiment analysis,” in *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020, pp. 2655–2666.
9. X. Zhou, Y. Wu, C. Zhang, and M. Huang, “Enhancing emotion classification with transfer learning and multi-task learning,” *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 1–16, 2022.