No Stupid Questions, Only Labeled Ones:

Intent Classification for University FAQs

Davide Pio Santangelo

Department of Interpretation and Translation

University of Bologna, Forlì

Abstract: As universities transition from static FAQ platforms to interactive conversational systems, intent classification of student queries emerges as a key challenge for effective human-machine communication. This paper explores how different modelling strategies perform in mapping queries to a structured set of seven intent categories, with a particular focus on the impact of synthetic augmentation. The models are trained on a newly created University FAQ dataset of manually collected questions, expanded with GPT-based augmentation. A classical baseline (TF-IDF + Linear SVM) is compared with neural architectures (BiLSTM and TextCNN), under both manual and augmented training regimes. Results show that augmentation consistently improves F1 and accuracy (+0.02 - 0.04) across all models. More complex architectures perform better but gain less from augmentation. Error analyses, comparative visualizations and an interactive demo offer practical insights. The study provides a robust foundation for building precise and adaptable intent classifiers in educational domains.

Key Words: Intent Classification, University FAQ, Data Augmentation, SVM, BiLSTM, TextCNN, Student Support, NLP

1. Introduction and Background

For university students, questioning is not only a path to knowledge but also a necessity of daily life. From admission and academic processes to housing, visas, and mobility information, international students face constant uncertainties and require real-time support. Universities have traditionally provided guidance by relying on digital helpdesks and FAQ web pages, but these systems are inherently limited and static: they force students to adapt their queries to rigid categories or submit request forms for later human follow-up.

To enable progress, many institutions are now experimenting with chatbots and virtual assistants, which can reduce staff workload and provide immediate, personalized assistance (Okonkwo & Abejide, 2021). However, the quality of these interactions depends critically on a system's ability to correctly recognize both explicit and implicit user intentions. While conversation feels effortless to humans because it is instinctive and deeply rooted in our evolution, its contextual richness and variability make it hard to codify for machines. Moreover, student queries can be phrased in countless ways and span a wide range of topics. Addressing these challenges is precisely the role of intent classification, a core task in natural language processing that maps unstructured questions to structured categories of meaning, identifying the underlying purpose behind a query. Accurate intent classification drives intelligent chatbots and conversational agents, enabling the transition from static platforms to more dynamic and interactive dialogue solutions (Pereira et al., 2023).

Two key factors drive this study. First, there is a scarcity of well-curated datasets and transparent pipelines that connect domain-specific data to clearly specified models, limiting systematic progress in educational NLP (Peyton et al., 2025; Pereira et al., 2023). Second, building large labeled datasets requires intensive manual effort, whereas artificial data generated by LLMs may offer a promising way to expand resources. Recent research suggests that GPT-based augmentation can indeed enrich models' training, though effects vary across contexts: while it has been shown to boost performance on intent classification tasks (Sahu et al., 2022; Madrueño et al., 2025), its benefits depend on intent clarity and often require human-in-the-loop validation. Similarly, structured augmentation pipelines have been found to systematically improve accuracy on educational datasets (Neshaei et al., 2025; Robson et al., 2021).

The goals are here both empirical and practical: to establish a clean and reproducible baseline for intent classification in the university FAQ domain, grounded in a well-documented dataset and experimental protocol; and to quantify the contribution of synthetic data on classification performance, analysing whether greater linguistic variability helps model generalization[1]. Accordingly, the work is structured around two research questions:

- **RQ1 (Modelling):** How do a strong classical baseline (linear SVM with TF-IDF) and lightweight neural models (BiLSTM and TextCNN with pretrained embeddings) compare on university FAQ intent classification?
- **RQ2 (Augmentation):** Does adding automatically generated questions to the training set help models perform better, or does it introduce noise that obscures intents?

The dataset consists of 3500 university-related questions, which were manually collected, annotated across seven intents, and later expanded with GPT-generated examples to a total of 5040 queries. While prior studies have demonstrated that even small datasets and simple classifiers can achieve strong results (Assayed et al., 2022), and that classical models often remain competitive baselines (Al-Tuama & Nasrawi, 2022), this project aims to systematically evaluate both classical and neural approaches on educational intent recognition. At the same time, the focus relies on comparing the two training regimes: manual-only vs. manual + synthetic.

The methodology section details the full experimental pipeline, while the results examine performance metrics across models and intents, with particular attention to where augmentation proves most beneficial. These findings are complemented by error analyses and comparative visualizations that offer deeper insights. Finally, an interactive demo illustrates the practical potential and usability of the models for real-world student queries. The paper concludes by discussing implications and future research.

2. Dataset and Project Foundations

The study builds on the first stage of the project, which focused on building and annotating a high-quality English-language dataset for university FAQ intent classification (Fig. 1). Motivated by the scarcity of structured and machine-readable resources for intent classification in academic contexts (Peyton et al., 2025; Okonkwo & Abejide, 2021), real or realistic student questions were collected from institutional websites worldwide. Using the BootCaT toolkit and a custom Python extraction script, the queries were automatically gathered and then manually reviewed to remove duplicates and out-of-domain entries, resulting in 2179 valid examples. To further enrich the dataset and capture greater diversity, 1321 additional questions were manually sourced from various university platforms and international FAQ pages. This combined effort led to the production of a manual dataset of 3500 questions, all related to the university ecosystem and student support.

A seven-intent taxonomy was meticulously developed to capture the most frequent student needs (Academic & Administrative; Accommodation & Housing; Admission & Application; Fees & Financial; Mobility & Exchange; Student Services & General Info; Visa & Legal Requirements). Annotation was carried out in INCEpTION under prescriptive guidelines, assigning a single intent label to each question. To ensure reliability, 20% of data was double-annotated, achieving almost perfect inter-annotator agreement (κ = 0.91). The finalized annotations were then exported and processed through a dedicated Python script to generate a clean and model-ready .xlsx file.

To increase linguistic variety and enable evaluation of synthetic augmentation, the dataset was further expanded with 1540 GPT-generated questions. A custom GPT assistant was created and trained with specific instructions to generate balanced sets of 220 queries per intent category. All outputs were then manually checked, and intent labels were corrected whenever the automatic intent assignment was incorrect. The resulting final corpus contains 5040 questions, structured in a clear and reproducible way, suitable for inspection and ready for subsequent model training. The present paper takes the next step: evaluating different modelling strategies and quantifying the impact of GPT-based augmentation on intent classification.

[1] Generalization in the field of natural language processing (NLP) is the ability of models to efficiently make predictions on previously unseen data based on what it has learned from the training data.

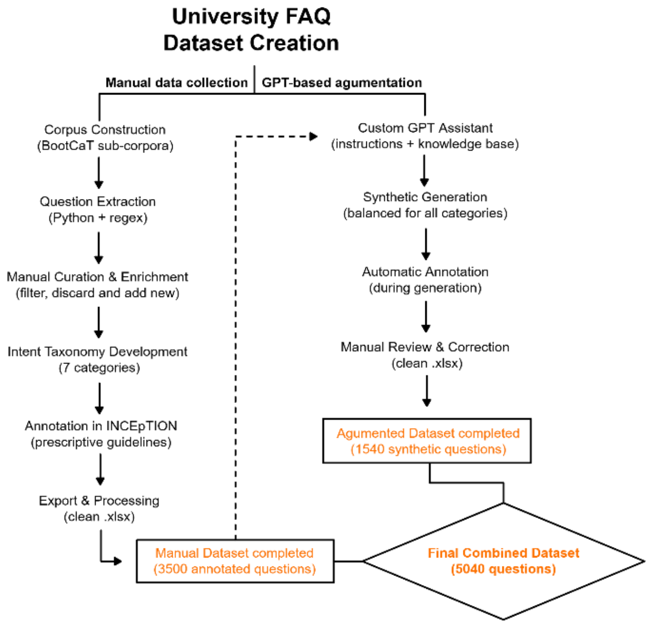


Fig. 1 – Dataset Creation and Augmentation Pipeline

3. Methodology and Experimental Setup

This section outlines the pipeline for training and evaluating intent classifiers on university-related queries, summarizing the code script[2] and covering preprocessing, data partitioning, feature representations, model architectures, training procedures and evaluation metrics. All choices aim at reproducibility, realistic evaluation, maximizing performance, and practicality for student-support deployment.

Intent classification is framed as a single-label classification task, with all experiments operating at the question (single-sentence) level. The objective is to automatically map each question to one of the 7 predefined intent categories covering key student-facing domains. Models are trained and compared under two training regimes designed to test the effect of synthetic augmentation:

- **Manual:** models are trained only on human-collected and annotated questions.
- **Augmented:** models are trained on the union of manual and GPT-generated questions.

Throughout the code and this paper, when a model is labeled "manual" or "augmented", this refers only to the training set composition.

3.1 General Data Preparation

Preprocessing: After uploading the dataset, the two data sources, manual and augmented, are loaded into separate data frames and checked for consistency in shape and scheme. Light text cleaning is applied, limited to whitespace trimming and lowercasing, while no stemming, lemmatization, or stopword removal is performed. This minimal normalization is intentionally adopted to preserve the natural linguistic variety of student queries and avoid removing semantic cues that could carry informative patterns or be decisive to detect intents.

Partitioning: Data splitting is performed only on the manual dataset, using stratification by intent to maintain label proportions. An 80/10/10 partition is chosen to maximize the amount of data available for training, given the relatively limited dataset size. A fixed random seed (42) ensures reproducibility. The GPT-based augmentation data is shuffled to avoid unwanted ordering patterns. It is then merged with the manual training set to form the augmented training regime, which is also re-shuffled to avoid ordering effects. The resulting splits maintain balanced intent proportions, with light imbalances being moderated after the addition of synthetic queries. In both training regimes, validation and test sets remain strictly manual-only. This design choice prevents synthetic paraphrases from leaking into evaluation, ensuring that test scores reflect real-world performance on genuine, human-written student queries. It also mirrors real deployment, where the system will encounter natural queries rather than synthetic ones.

Label Encoding: A single LabelEncoder from the Scikit-learn library is fitted on the intent labels found in the manual dataset. The encoder converts text-based intents into numeric IDs, producing a consistent mapping (label string → integer ID) that ensures reproducibility throughout the pipeline. This guarantees that the same intent is always assigned the same index across all experiments, training regimes and model comparisons.

[2] Portions of the code were developed with the support of ChatGPT (OpenAI), used strictly as an assistant to suggest possible solutions or approaches. All generated content was carefully reviewed and validated by the author, ensuring responsible and transparent use of AI tools.

3.2 Text representations

Two types of text representations are used, each tailored to the corresponding model family.

For the classical pipeline using SVM, queries are vectorized with TF–IDF, capturing both unigrams and bigrams (ngram_range=(1,2)), with a minimum document frequency (min_df=2) to drop very rare terms. Since each training regime has a different data distribution and uses features learned from its own training data, separate vectorizers are fitted per regime to keep the comparison fair. The shared manual validation and test sets are then transformed using the corresponding vectorizer.

Concerning the neural pipeline using BiLSTM and TextCNN, a Keras tokenizer is built for each training regime and fitted only on its corresponding training texts, because each regime has its own vocabulary and typical lengths. Queries are then converted into padded integer sequences, with vocabulary size and maximum sequence length computed independently for each type of training. The tokenizer uses an explicit "UNK" token to handle out-of-vocabulary words. Instead of padding sequences to the maximum length, the 95th percentile[3] is used as a cutoff (18 tokens for manual and 17 for augmented). This minimizes truncation and prevents inefficiently long padding driven by rare outlier questions, while preserving the vast majority of data.

3.3 Pre-trained embeddings

To provide the neural models with prior semantic knowledge and improve generalization, the embedding layer is initialized with pre-trained GloVe (Global Vectors for Word Representation) embeddings (6B 300-dimensional). GloVe is a set of pre-trained word vectors learned from large corpora (Wikipedia + Gigaword). It allows the model to start with meaningful representations of words rather than learning them from scratch on the FAQ dataset. For each training regime, an embedding matrix is built by aligning the tokenizer vocabulary with the GloVe vectors. Words without a pre-trained vector are zero-initialized and learned during training. The embedding layer remains trainable, allowing fine-tuning to the specific patterns of university-related queries, while still benefiting from the broad semantic knowledge encoded in GloVe. This is especially useful in a domain where user phrasing is highly variable.

3.4 Models

Three complementary modelling approaches are implemented for both training regimes: a solid classical baseline is compared with lightweight neural encoders that are widely used, computationally efficient and easy to reproduce. Neural model parameters and configurations were selected through repeated experimentation to maximize performance and reduce overfitting, while maintaining architectures comparable.

Linear SVM: A linear Support Vector Machine (LinearSVC) trained on TF–IDF features serves as a strong and cost-effective baseline for intent classification, offering stable and competitive performance. TF-IDF provides a simple yet effective representation of textual data as sparse numerical vectors, while Linear SVMs are known for their efficiency in intent detection and similar classification tasks. Scikit-learn default parameters are used to emphasize ease of replication.

BiLSTM: A bidirectional Long Short-Term Memory (BiLSTM) network is used to capture sequential dependencies and contextual information beyond fixed n-gram windows. Recurrent neural networks, and LSTMs in particular, can retain information across tokens, making them well-suited for intent detection where crucial cues may appear at any point in a query. The bidirectional variant is adopted to process the input text in both forward and backward directions, allowing the model to learn from past and future context simultaneously. The architecture consists of:

- Embedding layer initialised with GloVe embeddings (trainable);
- BiLSTM with 64 units and recurrent dropout (0.2) to prevent overfitting by limiting reliance on specific sequential patterns;
- Dropout layer (0.3) to further reduce overfitting;
- Dense layer with 32 ReLU units for non-linear feature combination;
- Softmax output layer over the seven intents, producing probability distributions.

[3] This is a common practice used to ignore extreme outliers while still capturing the majority of the training data. It covers 95% of typical sentence lengths and ignores the 5% longest outliers.

TextCNN: A Text Convolutional Neural Network (TextCNN) is the final approach. Convolutional architectures have proven highly effective for text classification because they can detect local n-gram patterns that are strong signals of intent. Unlike recurrent models that process text sequentially, CNNs process inputs in parallel, making them efficient with shorter texts such as FAQ questions. The convolutional filters, which are small matrices of weights, slide word embeddings to capture specific patterns of n-grams. The architecture includes:

- Embedding layer initialised with GloVe embeddings (trainable);
- Three parallel Conv1D branches with 64 filters and kernel sizes of 3, 4, 5 (capturing tri-, four- and five-gram patterns);
- GlobalMaxPooling1D on each branch to retain the most salient feature from each filter;
- Concatenation of pooled features from all branches;
- Dropout layer (0.3) to prevent overfitting;
- Dense layer with 32 ReLU units for non-linear feature combination;
- Softmax output layer over the seven intent categories.

Both neural models are trained using the Adam optimizer and sparse categorical crossentropy as the loss function, and accuracy is monitored during training. Architectural similarity is deliberately preserved to compare the impact of data augmentation rather than differences in model design.

3.5 Training procedure

Training settings are kept consistent across models to guarantee fair comparisons.

- Early stopping monitors validation loss with patience=3 and restore_best_weights=True. This guards against overfitting and reduces the effect of seed-specific fluctuations or “lucky runs”.
- Models are trained for a maximum of 15 epochs with a batch size of 32, which is sufficient for convergence under both manual and augmented regimes, given the use of pre-trained embeddings.
- To quantify run-to-run variability, each neural experiment is repeated across five random seeds (7, 42, 99, 123, 2025) and mean ± standard deviation results are reported. SVM remains deterministic under a fixed vectorizer.

This protocol balances rigor and practicality: early stopping combined with multiple seeds provides stable estimates, while extensive hyperparameter exploration was conducted to achieve strong performance and minimize overfitting.

3.6 Evaluation

Evaluation is conducted exclusively on the manual-only test set across all experiments. The validation set is used to guide early stopping and best model selection before making predictions on the test set. The following metrics are computed:

- Accuracy: overall percentage of correct predictions.
- Macro-F1: average F1 score giving equal weight to all intents, regardless of their size.
- Weighted-F1: average F1 score accounting for intent frequencies.
- Per-intent F1: individual F1 score for each intent label, useful for identifying harder or underperforming categories.

For all neural models, scores are reported as mean ± standard deviation across seeds. This provides a measure of variability and performance stability, rather than relying on a single run. In addition, results are displayed for the best run per model/regime (selected as the seed with the highest validation-based macro F1, but evaluated on the test set).

The manual and augmented regimes are compared for each model using summary tables and visualizations, such as F1 bar charts and accuracy boxplots. Cross-model comparisons are also provided, making it possible to identify which architectures benefit most from synthetic data and which ones perform better or make fewer misclassifications. Beyond standard performance scores, error analyses are carried out by comparing percentages of misclassified queries and generating confusion matrices and error lists. These diagnostics reveal where models fail, which intents are harder to detect, and whether augmentation helps by reducing systematic confusions rather than merely boosting performance numbers.

Finally, a Gradio-based interactive demo is included as a proof of concept, illustrating the practical applicability of the trained models. Users can freely select a model, type a university-related question, and receive a predicted intent in real time.

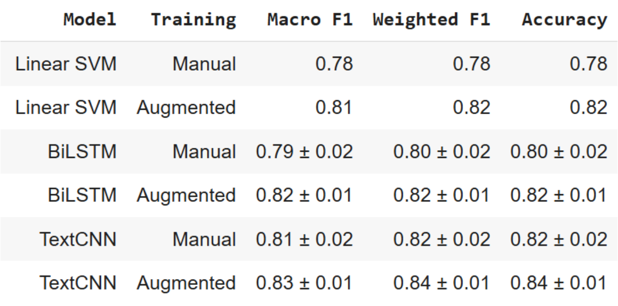
4. Results and Analysis

This section reports the main results, focusing on overall trends and cross-model comparisons.

Tables 1 and 2 together provide a comprehensive view of model performance across manual and augmented training regimes[4]. All three models perform competitively (Table 1), with Macro-F1, Weighted-F1 and Accuracy scores ranging between 0.78 and 0.84. Neural models outperform the Linear SVM, especially the augmented TextCNN, which achieves the best results (Macro-F1: 0.83; Weighted-F1: 0.84; Accuracy: 0.84); augmented neural models also show higher stability (± 0.01) compared to their manual counterparts.

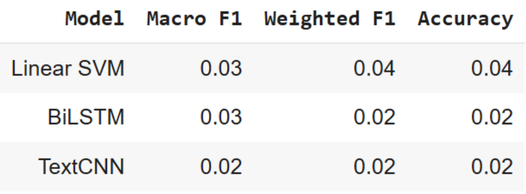
When focusing on the relative gains from augmentation (Table 2), the improvement trend becomes clearer. Augmentation consistently increases performance for every model, but the impact differs: Linear SVM benefits most (+0.03 - +0.04), while BiLSTM and TextCNN improve more modestly (+0.02 - +0.03). This indicates that augmentation is particularly helpful for simpler models, compensating for their limited generalization. By contrast, neural models already capture sequences and context effectively, so synthetic queries provide smaller improvements. Overall, these results show that GPT-based augmentation proves consistently beneficial, improving every architecture.

Table 1 – Global Summary of Performance across models and training regimes



Model	Training	Macro F1	Weighted F1	Accuracy
Linear SVM	Manual	0.78	0.78	0.78
Linear SVM	Augmented	0.81	0.82	0.82
BiLSTM	Manual	0.79 ± 0.02	0.80 ± 0.02	0.80 ± 0.02
BiLSTM	Augmented	0.82 ± 0.01	0.82 ± 0.01	0.82 ± 0.01
TextCNN	Manual	0.81 ± 0.02	0.82 ± 0.02	0.82 ± 0.02
TextCNN	Augmented	0.83 ± 0.01	0.84 ± 0.01	0.84 ± 0.01

Table 2 – Manual vs. Augmented Differences across models and training regimes



Model	Macro F1	Weighted F1	Accuracy
Linear SVM	0.03	0.04	0.04
BiLSTM	0.03	0.02	0.02
TextCNN	0.02	0.02	0.02

To complement the summary tables, Figures 2 and 3 provide a visual summary of the GPT augmentation effect on Macro F1 and Accuracy across models. In both scenarios, the orange bars (augmented) consistently exceed the yellow bars (manual), making the positive but modest improvements easier to appreciate at a glance. While the gains remain small in absolute terms, the side-by-side view confirms that synthetic data systematically benefits all three models.

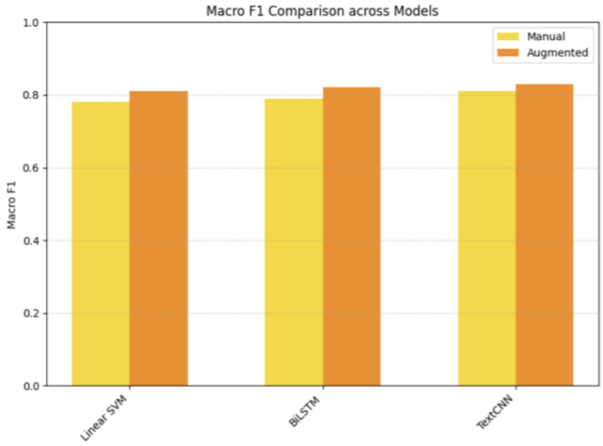


Fig. 2 – Bar Charts comparing Macro F1 across models and training regimes

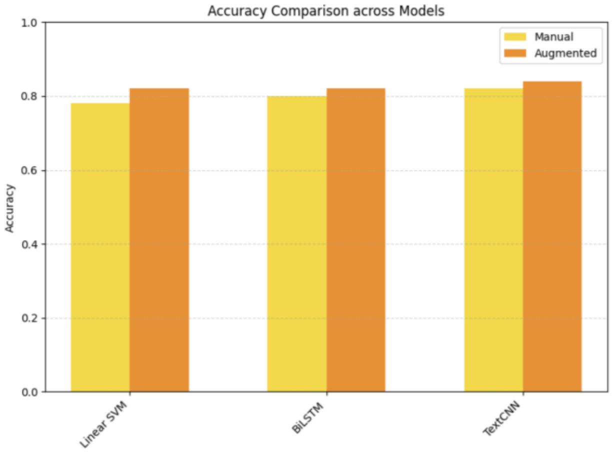


Fig. 3 – Bar Charts comparing Accuracy across models and training regimes

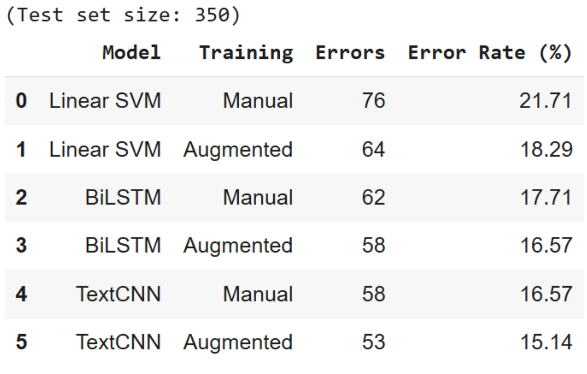
[4] Note: SVM results are deterministic single runs; BiLSTM and TextCNN results are reported as mean ± standard deviation across 5 random seeds.

Moreover, the best-performing runs for neural models confirm what has been observed, with TextCNN augmented reaching the highest scores:

- **BiLSTM – Manual (Seed 2025):** Macro F1 = 0.82, Weighted F1 = 0.82, Accuracy = 0.82
- **BiLSTM – Augmented (Seed 99):** Macro F1 = 0.84, Weighted F1 = 0.83, Accuracy = 0.83
- **TextCNN – Manual (Seed 2025):** Macro F1 = 0.83, Weighted F1 = 0.84, Accuracy = 0.83
- **TextCNN – Augmented (Seed 2025):** Macro F1 = 0.84, Weighted F1 = 0.85, Accuracy = 0.85

Table 4 reports the misclassification rates across models and training regimes. A clear pattern emerges: error counts decrease progressively from SVM to BiLSTM to TextCNN, with augmentation further lowering errors in every case. The Linear SVM, while showing notable improvement (from 76 to 64 errors), remains the weakest overall. Both neural models outperform SVM, with TextCNN trained on augmented data yielding the lowest error rate (15.1%), corresponding to just 53 misclassified queries out of 350. These findings reinforce that synthetic augmentation offers a measurable benefit.

Table 4 – Misclassified test set queries across models and training regimes



```
(Test set size: 350)
```

	Model	Training	Errors	Error Rate (%)
0	Linear SVM	Manual	76	21.71
1	Linear SVM	Augmented	64	18.29
2	BiLSTM	Manual	62	17.71
3	BiLSTM	Augmented	58	16.57
4	TextCNN	Manual	58	16.57
5	TextCNN	Augmented	53	15.14

Additional diagnostics confirm these trends:

- Confusion matrices show that augmentation benefits all models, with the largest gains for SVM, which improves across nearly every intent. In neural models, improvements are more selective, with occasional minor noise in categories that were already stable.
- Inspection of misclassified queries reveals that many errors are borderline or ambiguous, suggesting that models often fail where even human annotators might struggle.
- Per-intent F1 scores show that for SVM, augmentation improves performance across all intents, while for neural models, some intents improve, and the already strong ones remain stable.
- Variability analysis of Macro-F1 and Accuracy reveals that augmented models, particularly TextCNN, achieve slightly higher stability across seeds.

For more detailed error inspections with confusion matrices and misclassification examples, granular per-intent analyses, and distribution boxplots for variability, see Appendix A, B and C.

5. Conclusion and Future Research

This project investigated intent classification for university FAQ data, comparing a strong classical baseline with neural architectures and assessing the role of synthetic data augmentation.

Regarding the first research question, results showed that neural models outperformed the classical baseline. While the Linear SVM with TF-IDF provided a solid benchmark, both BiLSTM and TextCNN consistently achieved higher macro-F1 and accuracy, with TextCNN emerging as the most stable and effective. The error analysis further confirmed that neural models produced fewer misclassifications overall.

Addressing the second research question, synthetic data augmentation proved beneficial across all models (+0.02 - 0.04 on Macro F1 and Accuracy), though its impact varies. The SVM baseline gained the most, with augmentation acting as a compensatory mechanism that yielded noticeable improvements across nearly every intent. For neural models, gains were smaller but systematic, enhancing stability across seeds. Only minor noise was observed in a few intent categories, where error counts slightly increased. However, carefully curated augmentation can reliably enrich training data without harming generalization.

Nevertheless, limitations remain. The taxonomy was treated as a single-label problem, despite several queries could plausibly be multi-label; future work should investigate multi-label classification to better capture overlapping student needs. Moreover, the number of synthetic samples was kept uniform across intent categories, regardless of their specific size or difficulty; exploring targeted or intent-specific augmentation may therefore be a valuable direction. Finally, experiments were restricted to lightweight models for interpretability and reproducibility; future testing might involve more advanced architectures.

Overall, this study provides a reproducible benchmark that combines methodological insights with practical evidence for educational intent classification. By introducing a curated dataset and demonstrating the benefits of synthetic data augmentation, it lays the groundwork for future development of robust intent-based academic support systems.

References

Al-Tuama, Alaa T., and Dhamyaa A. Nasrawi. 2022. "Intent Classification Using Machine Learning Algorithms and Augmented Data." Paper presented at the 2022 International Conference on Data Science and Intelligent Computing (ICDSIC), Karbala University, Karbala, Iraq. IEEE.

Assayed, Suha K., Manar Alkhatib, and Khaled Shaalan. 2024. "Enhancing Student Services: Machine Learning Chatbot Intent Recognition for High School Inquiries." In BUiD Doctoral Research Conference 2023, edited by Khalid Al Marri, Farhan A. Mir, Susan A. David, and Mohamed Al-Emran, vol. 473 of Lecture Notes in Civil Engineering. Cham: Springer.

Assayed, Suha, Khaled Shaalan, and Manar Alkhatib. 2022. "A Chatbot Intent Classifier for Supporting High School Students." SSRN Scholarly Paper, Social Science Research Network.

Cutler, Ella; Zachary Levonian, and S. Thomas Christie. 2025. "Detecting Student Intent for Chat-Based Intelligent Tutoring Systems." arXiv.

Dinh, Hoa, and Thien Khai Tran. 2023. "EduChat: An AI-Based Chatbot for University-Related Information Using a Hybrid Approach." Applied Sciences 13, no. 22 (2023): Article 12446.

happyer. "Intent Recognition Technology." Medium. September 15, 2024. Last accessed September 18, 2025. https://medium.com/@threehappyer/intent-recognition-technology-e34962b2261b.

Label Your Data. "Intent Classification: Techniques for NLP Models." Label Your Data. Published July 30, 2025. Last accessed September 18, 2025. https://labelyourdata.com/articles/machine-learning/intent-classification.

Lyzr.ai. "Understanding Intent Recognition: Enhance User Interaction" Lyzr Glossaries. Last accessed September 18, 2025. https://www.lyzr.ai/glossaries/intent-recognition/#:~:text=Intent%20recognition%20is%20a%20pivotal,to%20understand%20and%20respond%20effectively.

Madrueño, Natalia, Alberto Fernández-Isabel, Rubén R. Fernández, Isaac Martín de Diego, and Gonzalo Polo Vera. 2025. "Exploring New Methods of Data Augmentation for Intent Classification Through Large Language Models." In Computational Science and Computational Intelligence. CSCI 2024, vol. 2501 of Communications in Computer and Information Science, 16-29. Cham: Springer.

Neshaei, Seyed Parsa, Richard Lee Davis, Paola Mejia-Domenzain, Tanya Nazaretsky, and Tanja Käser. 2025. "Bridging the Data Gap: Using LLMs to Augment Datasets for Text Classification." In Proceedings of the 2025 Educational Data Mining (EDM) Long Papers, paper 54. EDM.

Okonkwo, Chinedu Wilfred, and Abejide Ade-Ibijola. 2021. "Chatbots Applications in Education: A Systematic Review". Computers and Education: Artificial Intelligence, vol. 2, no. 2, 2021, p. 100033.

Pereira, D. S. M., Falcão, F., Costa, L., Lunn, B. S., Pêgo, J. M., & Costa, P. 2023. "Here's to the future: Conversational agents in higher education - a scoping review". International Journal of Educational Research,122, 102233.

Peyton, Kevin, Saritha Unnikrishnan, and Brian Mulligan. 2025. "A Review of University Chatbots for Student Support: FAQs and Beyond." Discover Education 4 (2025): Article 21.

Robson, Paula, Aguiar Neto, D., Romero, D., & Guerra, P. 2021. "Evaluation of Synthetic Datasets Generation for Intent Classification Tasks in Portuguese". Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, (pp. 265-274).

Sahu, Gaurav; Pau Rodriguez, Issam H. Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. "Data Augmentation for Intent Classification with Off-the-Shelf Large Language Models." arXiv.

Sapardic, Jelisaveta. "What Are Chatbot Intents: Classification, Use Cases, and Training Tips." Tidio Blog. April 15, 2025. Last accessed

September 18, 2025. https://www.tidio.com/blog/chatbot-intents/.

Sayedi, Husna. “Intent Recognition in NLP.” TAUS Resources Blog. September 7, 2021. Last accessed September 18, 2025. https://www.taus.net/resources/blog/intent-recognition-in-nlp.

Tapereal.com. “Chatbot Intent Classification Guide 2024.” Tapereal Blog. Last accessed September 18, 2025. https://web.tapereal.com/blog/chatbot-intent-classification-guide-2024/.

Appendix

A) Confusion matrices and Error examples

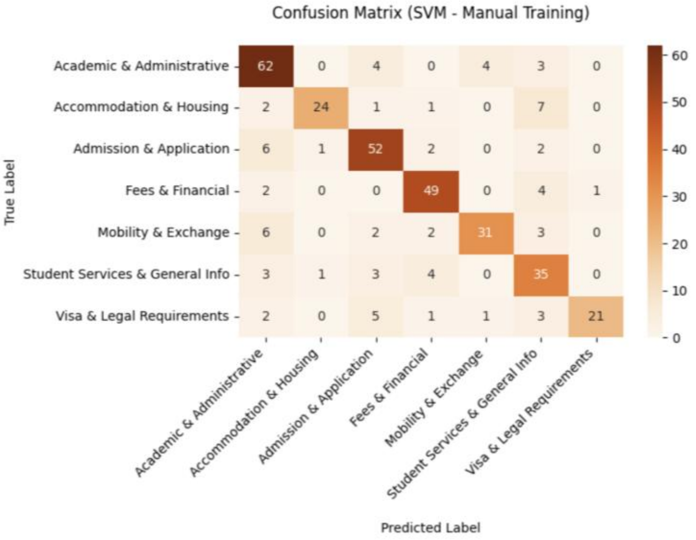
To further inspect model behaviour, confusion matrices are computed for the SVM and for the best runs of each neural architecture. These visualizations highlight per-intent strengths and weaknesses. Augmentation improves performance across all models, though with different effects.

For SVM, it consistently reduces misclassifications across almost all intent categories, most visibly in “Mobility & Exchange” (13 → 10) and “Admission & Application (11 → 8).

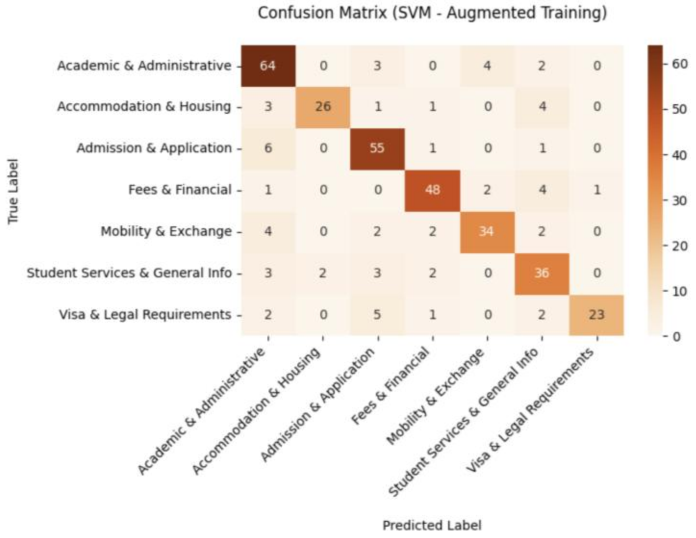
For BiLSTM, the picture is more mixed: “Accomodation & Housing” (6 → 1) and “Visa & Legal Requirements” (13 → 10) improve substantially, while errors increase in “Fees & Financial” (5 → 8) and “Student Services & General Info” (5 → 7). This shows that augmentation sometimes may introduce noise in already stable categories.

TextCNN displays the sharpest diagonal overall, but augmentation has uneven effects. It produces large improvements in “Admission & Application” (13 → 6) and “Mobility & Exchange” (7 → 4), while degrading “Student Services & General Info” (7 → 9) and “Accomodation & Housing” (3 → 5).

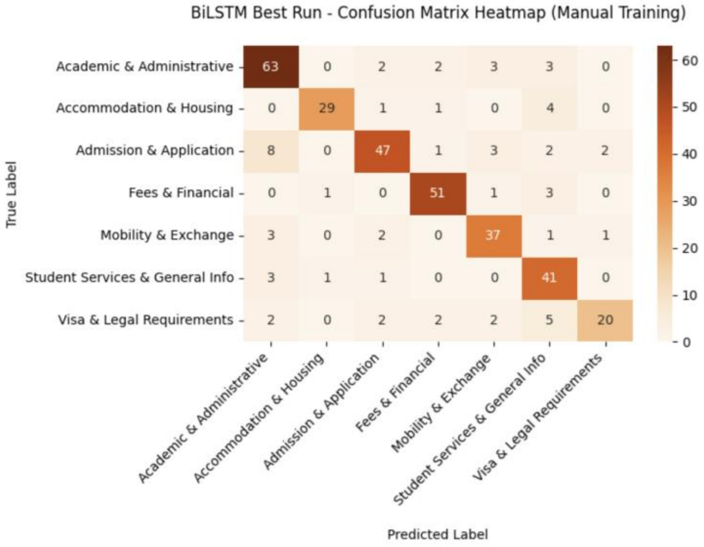
In general, all architectures successfully learnt the intent structure. Augmentation improves SVM performance the most, supporting the observation that synthetic data is especially beneficial for the weaker baseline. Diagonals are strongest for “Fees & Financial” and “Academic & Administrative”, while “Visa & Legal Requirements” remains the most error-prone, reflecting its smaller category size and less definite boundaries.


Confusion Matrix (SVM - Manual Training)

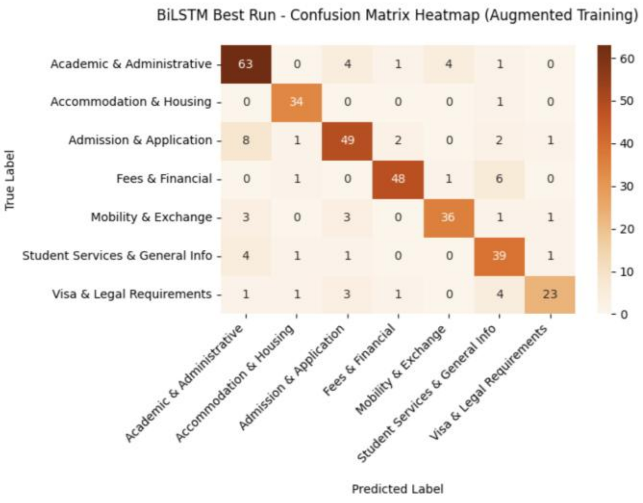
True Label \ Predicted Label	Academic & Administrative	Accommodation & Housing	Admission & Application	Fees & Financial	Mobility & Exchange	Student Services & General Info	Visa & Legal Requirements
Academic & Administrative	62	0	4	0	4	3	0
Accommodation & Housing	2	24	1	1	0	7	0
Admission & Application	6	1	52	2	0	2	0
Fees & Financial	2	0	0	49	0	4	1
Mobility & Exchange	6	0	2	2	31	3	0
Student Services & General Info	3	1	3	4	0	35	0
Visa & Legal Requirements	2	0	5	1	1	3	21

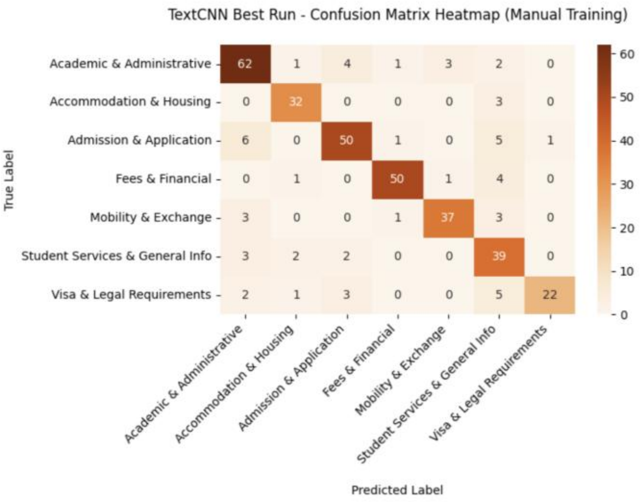

Confusion Matrix (SVM - Augmented Training)

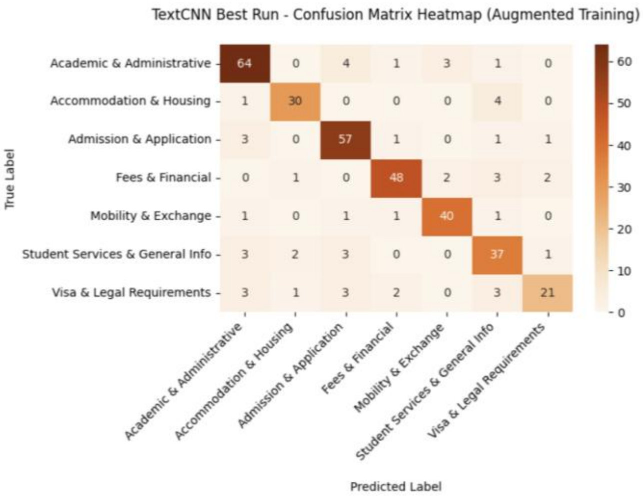
True Label \ Predicted Label	Academic & Administrative	Accommodation & Housing	Admission & Application	Fees & Financial	Mobility & Exchange	Student Services & General Info	Visa & Legal Requirements
Academic & Administrative	64	0	3	0	4	2	0
Accommodation & Housing	3	26	1	1	0	4	0
Admission & Application	6	0	55	1	0	1	0
Fees & Financial	1	0	0	48	2	4	1
Mobility & Exchange	4	0	2	2	34	2	0
Student Services & General Info	3	2	3	2	0	36	0
Visa & Legal Requirements	2	0	5	1	0	2	23


BiLSTM Best Run - Confusion Matrix Heatmap (Manual Training)

True Label \ Predicted Label	Academic & Administrative	Accommodation & Housing	Admission & Application	Fees & Financial	Mobility & Exchange	Student Services & General Info	Visa & Legal Requirements
Academic & Administrative	63	0	2	2	3	3	0
Accommodation & Housing	0	29	1	1	0	4	0
Admission & Application	8	0	47	1	3	2	2
Fees & Financial	0	1	0	51	1	3	0
Mobility & Exchange	3	0	2	0	37	1	1
Student Services & General Info	3	1	1	0	0	41	0
Visa & Legal Requirements	2	0	2	2	2	5	20







Here are a few illustrative misclassification examples reported per model and regime, highlighting different types of errors (e.g., semantic overlap, ambiguity, and borderline cases).

SVM Manual Training

- *"do i have to bring a computer?"*
  True: Student Services & General Info →
  Predicted: Accommodation & Housing
  (personal logistics vs. housing needs).

- *"when is the deadline to provide my legalised documents?"*
  True: Visa & Legal Requirements →
  Predicted: Admission & Application
  (confusion between administrative and legal paperwork).

- *"when do i get instructions for the erasmus+ grant?"*
  True: Mobility & Exchange →
  Predicted: Fees & Financial
  (confusion because of the word "grant").

SVM Augmented Training

- *"what are the main deadlines to be respected?"*
  True: Student Services & General Info →
  Predicted: Admission & Application
  (deadlines can be both administrative or service-related, reflecting ambiguity).

- *"how do i view my assignment/roommates and change them?"*
  True: Accommodation & Housing →
  Predicted: Academic & Administrative
  (confusion between housing allocation vs. academic course assignment).

- *"does credit recognition cost anything?"*
  True: Fees & Financial →
  Predicted: Academic & Administrative
  (financial vs. administrative framing of credit recognition).

BiLSTM Manual Training

- *"what is the deadline for uploading my thesis?"*
  True: Academic & Administrative →
  Predicted: Admission & Application
  (deadlines can be either program requirements or admission processes).

- *"how can i appeal my accommodations decision?"*
  True: Student Services & General Info →
  Predicted: Accommodation & Housing
  ("accommodations" misread as housing rather than support services).

- *"what immigration documents will i need in order to travel?"*
  True: Visa & Legal Requirements →
  Predicted: Mobility & Exchange
  (confusion between travel logistics and legal entry requirements).

BiLSTM Augmented Training

- *"what happens if i don't have these documents?"*
  True: Admission & Application →
  Predicted: Student Services & General Info
  (uncertainty about paperwork classified as generic student support rather than admissions).

- *"are there any travel allowances?"*
  True: Fees & Financial →
  Predicted: Student Services & General Info
  (funding vs. general student benefits confusion).

- *"how long can i be a visiting student?"*
  True: Academic & Administrative →
  Predicted: Student Services & General Info
  (borderline case: administrative rule vs. general information).

TextCNN Manual Training

- *"what is the port of entry and what needs to be done?"*
  True: Visa & Legal Requirements →
  Predicted: Admission & Application
  (travel/visa logistics mistaken for application paperwork).

- *"can you send me my degree certificate?"*
  True: Academic & Administrative →
  Predicted: Student Services & General Info
  (similarity between academic document request vs. support service query)

- *"when do i receive the travel support?"*
  True: Mobility & Exchange →
  Predicted: Student Services & General Info
  (exchange-related aid confused due to the absence of an Erasmus-related term)

TextCNN Augmented Training

- *"how do i pay for my tuition while on exchange?"*
  True: Fees & Financial →
  Predicted: Mobility & Exchange
  (financial matters blurred with exchange program logistics).

- *"do you provide homestay services?"*
  True: Accommodation & Housing →
  Predicted: Student Services & General Info
  (housing-related query softened into generic support services).

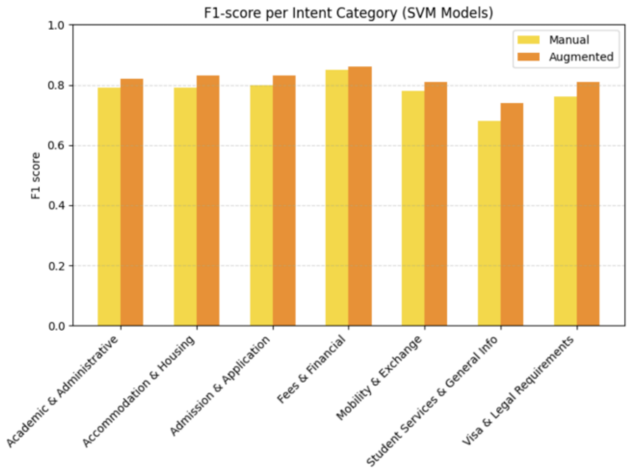
- *"how do i check my dining dollar balance?"*
  True: Fees & Financial →
  Predicted: Student Services & General Info
  (financial vs. general campus services).

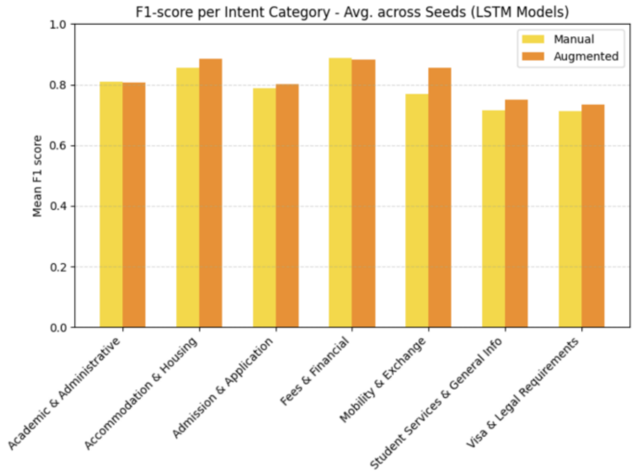
Many misclassifications appear borderline rather than absolute errors, reflecting the inherent ambiguity of some student questions. In such cases, where even human annotators might disagree in choosing the intent, a multi-label setup could arguably be more appropriate. This observation suggests that the models are not failing dramatically but rather struggling with the same fuzzy boundaries that exist in real-world intent classification.

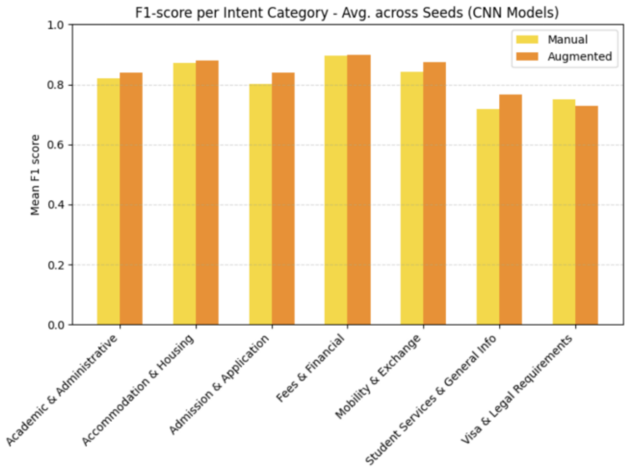
B) Per-intent F1 scores

Three bar charts examine per-intent F1 scores across models and training regimes. These visualizations highlight which specific intent categories benefit most from synthetic data.

Across all three graphs, augmentation generally shifts bars upward, indicating consistent per-intent F1 gains. For the SVM, the lift is visible on nearly every category, with the most noticeable jumps on the last two ("Student Services & General Info" and "Visa & Legal Requirements"). For BiLSTM, improvements are again broad-based: "Mobility & Exchange" shows clear gains, while already strong intents, like "Fees & Financial", remain high. For TextCNN, synthetic data mostly improves or preserves performance, but shows a slight dip for "Visa & Legal Requirements". Overall, the improvements after augmentation are visible, but small differences between the intents suggest that intent-specific augmentation may provide further benefit.

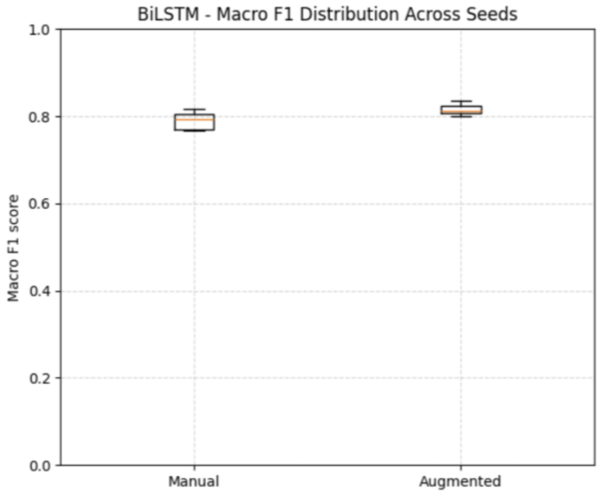


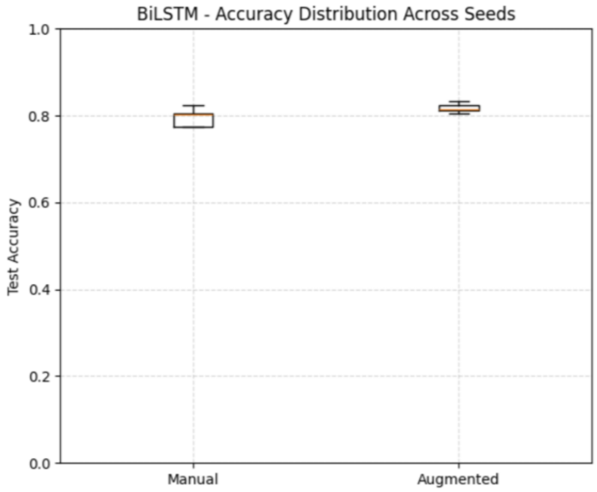


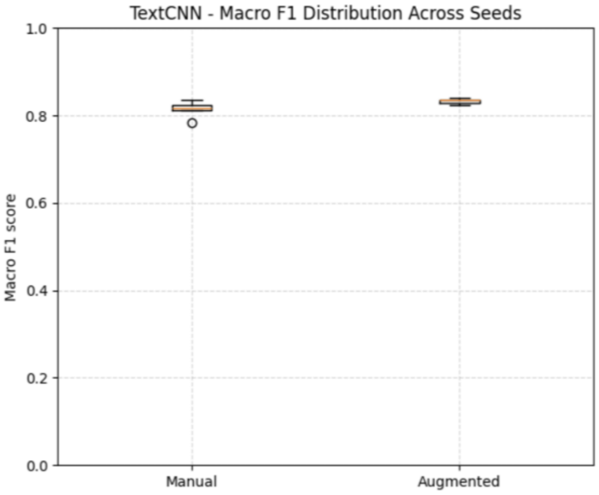


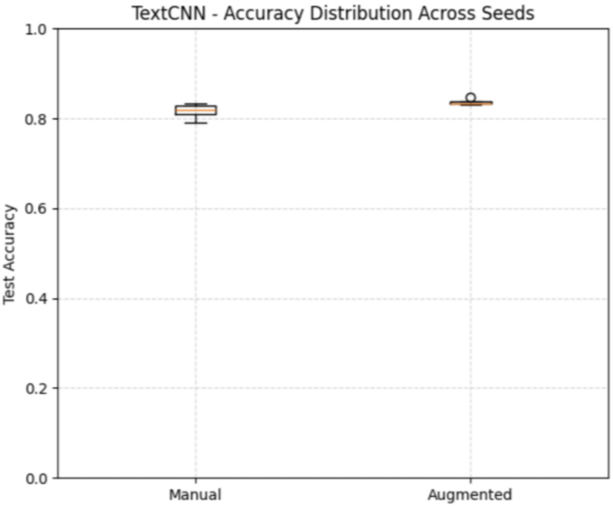
C) Macro F1 and Accuracy Variability

Four boxplots, two for each neural model, are generated to visualize how Macro-F1 and Accuracy vary across five random seeds. In all plots, the median (centre orange line) for the augmented regime is higher, and the IQR/whiskers are slightly tighter. This confirms that augmentation yields small but consistent gains and equal or better stability. For TextCNN in particular, the augmented distributions are notably tight, indicating very high stability. Overall, aligning with the earlier visualizations, gains are modest but systematic.






TextCNN - Macro F1 Distribution Across Seeds
Macro F1 score
1.0
0.8
0.6
0.4
0.2
0.0
Manual
Augmented


TextCNN - Accuracy Distribution Across Seeds
Test Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Manual
Augmented