Intent Classification using ATIS Dataset for Natural Language Understanding

Anbu Ezhilmathi Nambi

School of Engineering and Sciences, George Washington University
Washington, DC, USA

Abstract

Intent Classification is a fundamental task in Natural Language Understanding that enables systems to interpret and categorize user queries effectively. This paper investigates the performance of various machine learning techniques for intent classification using the Airline Travel Information Systems (ATIS) dataset. Traditional approaches like Multinomial Naïve Bayes and Logistic Regression, are compared with advanced techniques like Convolutional Neural Networks (CNNs) with pre-trained GloVe embeddings and BERT model. The models are evaluated using accuracy and F1 score as metrics. The results show a performance hierarchy, with BERT achieving the highest accuracy of 95.4% and an F-1 score of 94.6%. These findings highlight the superiority of transformer-based models for intent classification and their potential to enhance travel-related virtual assistants.

1 Introduction

Intent classification is a key task in Natural Language Understanding, where the goal is to accurately identify the purpose behind a user's query. In airline-related virtual applications intent classification plays an important role in ensuring that user requests such as booking flights, checking flight statuses, or inquiring about ground services are addressed effectively. With the growing reliance on virtual assistants in travel and other industries, achieving high accuracy in understanding user queries remains challenging due to the variability in language and context of the intent classification is vital for enhancing user services efficiently.

To tackle this problem, a range of machine-learning approaches is applied to the ATIS dataset, a widely used benchmark for intent classification in airline services. For traditional machine learning methods like Multinomial Naïve Bayes and Logistic Regression, pre-processing steps such as tokenization, stopwords removal, TF-IDF vectorization, and label encoding are employed to prepare the data. Advanced methods, including Convolutional Neural Networks (CNNs) and Bidirectional encoder representations from transformers (BERT), incorporate pre-trained GloVe embeddings to capture the semantic nuances and context in the text more effectively.

The findings reveal a clear performance hierarchy among these models. Traditional approaches like Logistic Regression and Naïve Bayes provide a solid baseline, while the CNN model significantly outperforms them, with BERT achieving the highest accuracy (95.4%) and F1 score (94.6%). These findings underscore the transformative potential of transformer-based architectures in improving intent classification, particularly in the context of airline-related queries.

2 Methodology

2.1 Dataset Overview

The Airline Travel Information System (ATIS) dataset is a widely used benchmark for intent classification in natural language processing (NLP). It consists of user queries related to airline travel, such as flight bookings, fare inquiries, schedule requests, and baggage tracking. Each query is labeled with one of 17 intent classes, representing diverse user requests. These intent classes capture various aspects of airline travel, ranging from flight details to specific service requirements.

Researchers collected the dataset from real-world user queries that were submitted to airline booking systems or customer service bots. These queries

were manually labeled, providing valuable resources for training and evaluating intent classification models. The ATIS dataset is available on the HuggingFace platform, that provides access to variety of NLP datasets.

The dataset is divided into three subsets:

• **Training set:** 4,375 samples used to train models.

• **Validation set:** 597 samples used to tune models and evaluate intermediate performance.

• **Test set:** 895 samples used for final evaluation and model comparison.

2.2 Class Distribution

The ATIS dataset exhibits an imbalance distribution across the intent classes. The "Flight" intent is the most frequent class, while intent like "Meal" and "Restriction" are much rarer. This class imbalance presents challenges during model training as models become biased toward the majority class, leading to lower performance in minority classes. As illustrated in the table below:

Intent Class	Count
Flight	3,226
Airfare	372
Ground service	224
Airline	138
Abbreviation	129
Aircraft	71
Flight time	48
Quantity	45
Airport	18
Distance	18
Flight and airfare	18
City	17
Ground fare	16
Capacity	14
Flight no	11
Meal	5
Restriction	5

Table 1: Intent Class Counts

2.3 Data Preprocessing

The data is preprocessed and prepared before feeding into the models, this process ensures the text data is clean and suitable for machine learning models. These steps are outlined as follows:

Text Tokenization: Each sentence in the dataset was split into individual words or tokens. This step is crucial for converting text into numerical representation that machine learning models can process.

Text Normalization: Text normalization is the process of cleaning the data by removing unnecessary elements. The text normalization techniques include:

- **Lowercasing:** All text was converted to lowercase to maintain consistency and to prevent the model from treating the same word in different cases as different entities.
- **Stop Words Removal:** Common but uninformative words (e.g., the, a, and) were removed from the text to reduce noise and focus on meaningful content.

Label Encoding: The 17 intent labels in the dataset are categorical, and label encoding was applied to convert these labels into numerical format. This step is essential for feeding the labels into machine learning models, as they expect numerical input for classification tasks.

Padding: Padding was applied to ensure all input sentences were the same length. This is important for models like CNN and BERT, as they require fixed-length input sequences for efficient training.

2.4 Model Selection

In this study, both traditional machine learning models and deep learning models are utilized to classify the intents in the dataset. The traditional machine learning models are used as the baseline models for this study.

Baseline Models

- **Multinomial Naïve Bayes**: A probabilistic classifier model based on Baye's theorem to predict class probabilities. It assumes that the features in this case words or tokens are conditionally independent given the class label. This model is commonly used for text classification tasks due to its ability to handle frequency-based data.
- **Logistic Regression**: A linear model for binary and multi-class classification was applied with a one-vs-rest strategy. The model builds separate binary classifiers for each class and selects the one with the highest confidence.

While these models are relatively simple, they provide a good starting point for comparison against more complex models.

Convolutional Neural Networks (CNNs):

Convolutional Neural Networks (CNNs) are a specialized type of neural network designed to extract hierarchical features from data. While CNNs are used for deep learning, they have also demonstrated significant effectiveness for text data by learning local patterns in a sequence of words. In the context of intent classification, CNNs can identify contextual patterns in utterances, such as specific keywords or phrase-level features, that are indicative of an intent.

CNN Model Design

The architecture of the CNN model for intent classification included the following components:

- **Embedding layer:** This layer was initialized with the GloVe embedding matrix and was frozen during training to leverage the pre-trained semantic features.
- **Convolutional Layer:** A Conv1D layer with 32 filters and a kernel size of 8 was used to capture local patterns in sequences.
- **MaxPooling Layers:** This layer was used to reduce the dimensionality of the feature maps while retaining the most salient features ensuring computational efficiency.
- **Fully Connected Layers:** This is a dense hidden layer with 10 units and ReLU activation captured complex interactions between featured, followed by an output layer with softmax activation for multiclass classification.

Bidirectional encoder representations from transformers (BERT):

BERT-Base-Cased is a transformer-based model pre-trained on a large corpus, including Books Corpus and English Wikipedia. Unlike traditional models that process text unidirectionally, **BERT-Base-Cased** utilizes bidirectional context to capture the relationships between words in a sentence. This bidirectional approach enables BERT to understand both the left and right context of each word, providing a more comprehensive understanding of sentence meaning. In the context of intent classification, this enhanced contextual understanding allows **BERT-Base-Cased** to more accurately predict user intents by effectively capturing the nuanced relationships between words.

BERT Fine-tuning:

The BERT model used in this study was fine-tuned for intent classification. The architecture included:

- **BERT Encoder** extracts contextual embedding from input texts. The encoder processes the text in a bidirectional manner, ensuring that the relationships between words are captured comprehensively.
- **Classification Layer** is a fully connected layer with softmax activation that is added on top of the BERT encoder for multiclass classification. This layer maps the contextual embedding to the final class probabilities for intent prediction.
- **Fine-tuning** was performed using AdamW optimizer with a learning rate of 1e-5 and a liner learning rate scheduler. Fine-tuning adjusts the pre-trained weights of BERT to better align with the specific intent classification task.

2.5 Model Comparisons

The performance of traditional and advanced machine learning models was compared to identify the best approach for intent classification.

Both CNN and BERT were used to classify intents, with BERT outperforming CNN due to its ability to capture bidirectional relationships and deeper semantic understanding.

While CNNs are effective at learning local features, their sequential nature and lack of pre-trained context limit performance compared to transfer-based models like BERT.

3 Experiments

The models are trained and evaluated; they are used to make predictions on new, unseen queries. The process involves tokenization and normalization of the input text. Then the input text is passed through the models to predict the intent. The predicted intents were then decoded from numerical labels back to their corresponding class names. The models' prediction was analyzed to assess their accuracy in real-world scenarios, ensuring that they could generalize well to new, unseen queries.

3.1 Model Training

3.1.1 Baseline models

Two baseline machine learning models Multinomial Naive Bayes and Logistic Regression were utilized for intent classification. These models were trained on TF-IDF features that was extracted from the preprocessed dataset. The Naive Bayes model was fitted to the training data using the fit() method, while Logistic Regression was configured to a maximum of 1000 iterations to ensure convergence. After training, both models were evaluated on the test set, and their performance was thoroughly assessed.

3.1.2 CNN

The model was compiled using Adam optimizer and categorical cross-entropy loss. It was trained for 10 epochs with a batch size of 32. By employing this CNN-based approach, the model effectively learned to classify intents based on patterns in utterances, demonstrating the suitability of CNNs for this task.

3.1.3 BERT

The model was trained for six epochs using a batch size of 16. By fine-tuning BERT for intent classification, the model leveraged its pre-trained language knowledge while tailoring the model's understanding to the specific task, resulting in improved performance in intent prediction.

3.2 Evaluation and Performance Metrics

To evaluate the performance of the models, two key metrics are used:

Accuracy: This metric measures the percentage of correct predictions out of all predictions made. It is a standard metric used in classification tasks.

F1 Score: This measures the harmonic mean of precision and recall, providing a balanced measure of a model's ability to correctly classify positive examples while minimizing false positives and false negatives. This metric is especially useful when the dataset has imbalanced classes.

Both metrics were calculated for each model on the test set to provide an objective comparison of their performance.

4 Results

The results of the experiments are summarized in the table below, which shows accuracy and F1 scores for each model on the test set:

Models	Accuracy	F1-Score
Naïve Bayes	76.3%	74.5%
Logistic Regression	82.1%	80.8%
CNN	90.2%	89.1%
BERT (Fine-tuned)	95.4%	94.6%

Table 2: Models Performance Comparison

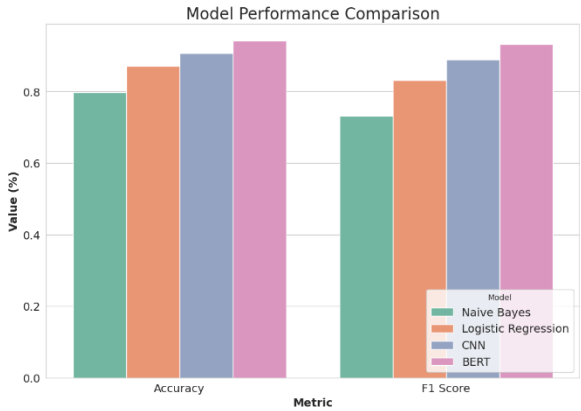

Figure 1: Graph of Model Performance Comparison

BERT outperformed all other models achieving the highest accuracy and F1-score. Among baseline models, Logistic Regression performed better than Multinomial Naïve Bayes. CNN showed significant improvements over traditional methods but fell short of BERT's performance.

Baseline Predictions
Naïve Bayes:
Utterance: determine the type of aircraft used on a flight from cleveland to dallas that leaves before noon
Predicted Class: flight
Actual Class: aircraft
Logistic Regression:
Utterance: determine the type of aircraft used on a flight from cleveland to dallas that leaves before noon
Predicted Class: aircraft
Actual Class: aircraft

Table 3: Baseline Model's Predictions Comparison on Test data

5 Analysis

5.1 Baseline models

Multinomial Naïve Bayes struggles with complex patterns in the dataset, this is likely due to its assumption of feature independence.

Naïve Bayes Prediction on Input
Input: how far is toronto international from downtown **Predicted Class:** flight **Actual Class:** distance

Table 4: Naïve Bayes Model's Prediction on unseen data

Logistic Regression demonstrated good performance on the test data, effectively leveraging the TF-IDF vectorized inputs to make accurate predictions, as shown in Table 3. However, the model struggled to generalize to user-provided queries, often failing to accurately classify intent that were phrased differently or contained unseen patterns as illustrated in Table 5 below.

Logistic Regression Prediction on Input
Input: how far is toronto international from downtown **Predicted Class:** flight **Actual Class:** distance

Table 5: Logistic Regression Model's Prediction on unseen data

Both model's predictions are inaccurate when put to test on some inputs of different classes, possibly because the "Flight" intent is the most common, while other intents are significantly less frequent. This imbalance poses challenges in model training, leading to a bias toward the majority class.

5.2 Advance Models

CNN demonstrates the ability to capture hierarchical features from text, resulting in significantly improved performance over traditional models.

CNN Prediction on Input
Input: how far is toronto international from downtown **Predicted Class:** distance **Actual Class:** distance

Table 6: CNN Model's Prediction on unseen data

BERT excelled due to its ability to understand context through its bidirectional transformer architecture. Its pre-trained knowledge provided a strong foundation and fine-tuned further optimized its performance for the dataset.

BERT Prediction on Input
Input: how far is toronto international from downtown **Predicted Class:** distance **Actual Class:** distance

Table 7: BERT Model's Prediction on unseen data

5.3 Class-wise Analysis

Both CNN and BERT demonstrated strong performance in predicting the intent classes for unseen data, as evidenced by Tables 6 and 7.

These models successfully identified the majority of the classed, showing their ability to generalize well.

However, as shown in Table 8, minor classes such as "meal" and "restriction" presented challenges due to their small sample sizes. Despite the overall strong performance, both models struggled with less frequent classes, highlighting the impact of class imbalance on the model's ability to recognize minor intent classes.

CNN
Input: what meals are served on american flight 811 from tampa to milwaukee **Predicted Class:** flight **Actual Class:** meals
Input: are there any restrictions for children traveling alone on united airlines **Predicted Class:** abbreviation **Actual Class:** restriction
BERT
Input: what meals are served on american flight 811 from tampa to milwaukee **Predicted Class:** airfare **Actual Class:** meals
Input: are there any restrictions for children traveling alone on united airlines **Predicted Class:** abbreviation **Actual Class:** restriction

Table 8: CNN & BERT Model's Predictions on Minor Classes

5.4 Confusion Matrix

5.4.1 Baseline Models

Both baseline models, Naive Bayes and Logistic Regression, provide a foundation for intent classification but show limitations in accurately distinguishing certain intents, particularly patterns like “ground_service” and “city”. While Naive Bayes and Logistic regression offer decent performance for some intents, both models are significantly outperformed by CNN and BERT which are explained in detail.

5.4.2 CNN

CNN model's confusion matrix demonstrates a strong diagonal, indicating high accuracy in classifying most intents, particularly “flight”, “airfare”, and “abbreviation”, highlighting clear patterns within the data for these intents. Although there remains some confusion between similar intents such as "ground_service" and "city", CNN's capacity to identify intricate patterns enhances generalization and minimizes misclassifications.

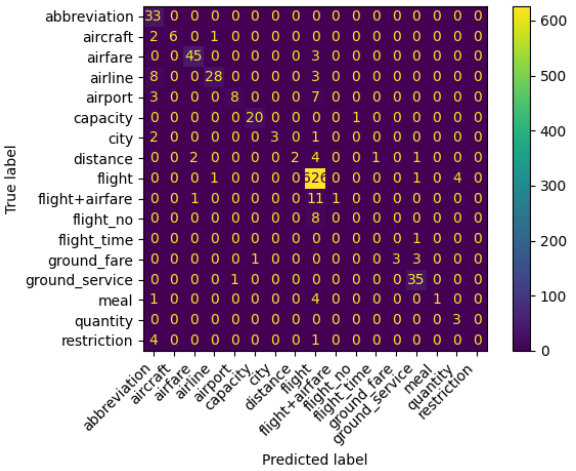

Figure 2: Confusion Matrix (CNN)

5.4.3 BERT

The BERT model's confusion matrix showcases exceptional performance in intent classification, significantly outperforming all other models. Its dominant diagonal reflects high accuracy across most intents, indicating BERT's ability to effectively capture the distinctive features of each intent. The overall clarity and minimal off-diagonal elements highlight BERT's robust generalization and accurate differentiation of intent categories.

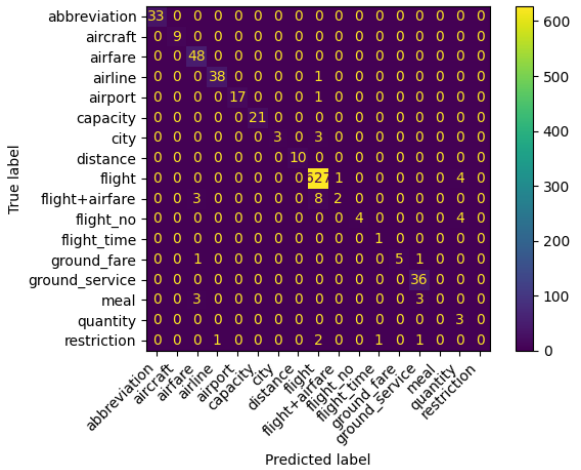

Figure 3: Confusion Matrix (BERT)

Analysis of the confusion matrix revealed that most misclassifications occurred between semantically similar intent, such as “ground service” and “ground fare.” This highlights the need for more robust representations for these overlapping intents.

The darker diagonal elements in their confusion matrices imply that the CNN and BERT models perform well on the dominating intent classes. But in general, BERT is more accurate and manages less frequent classes better, which results in fewer off-diagonal mistakes. This implies that BERT can identify more intricate connections in the data. CNN performs rather well, although it frequently misclassifies some intent classes, particularly ones with fewer training examples. Essentially, BERT has an advantage in intent classification due to its transformer architecture's contextual awareness, which produces a confusion matrix that shows improved accuracy and intent differentiation.

6 Discussions

The results of this study emphasize the significant advancements that transformer-based models, particularly BERT, bring to intent classification tasks. BERT's capability to understand nuanced relationships in language through bidirectional contextual embeddings underscores its superior performance compared to traditional machine learning models and CNNs. The model’s accuracy and F1 score demonstrate its effectiveness in practical scenarios, such as virtual assistant systems in the airline domain.

However, these benefits come with challenges. The higher computational demands of BERT, including

memory and processing time, make it less accessible for resource-constrained environments. Additionally, the class imbalance in the ATIS dataset negatively impacted predictions for minority classes, indicating a need for strategies like oversampling, under-sampling, data augmentation, or fine-tuned loss functions to address these gaps in the future.

From a deployment perspective, while models like Logistic Regression offer simplicity and speed, their accuracy is insufficient for high-stakes applications. This highlights the trade-off between complexity and performance that must be considered depending on the intended use case.

7 Conclusion and Future Works

7.1 Conclusion

This research demonstrates the superiority of advanced language models like BERT in intent classification tasks. The findings show that leveraging pre-trained embeddings allows for a deeper understanding of context, enabling significant improvements over traditional methods. The study also highlights the importance of addressing class imbalance and computational constraints for broader applicability. Overall, this study underscores the transformative potential of modern NLP techniques for intent classification, paving the way for more accurate, reliable, and practical natural language understanding applications.

7.2 Future Work

Future work could build on this foundation by implementing data augmentation and leveraging transfer learning to address the challenge of underrepresented classes. Data augmentation techniques, such as generating synthetic examples through synonym replacement or back-translation, can increase the diversity and size of minority class samples, thereby improving model generalization and reducing class imbalance issues. Transfer learning, such as fine-tuning transformer models on targeted tasks, can help the model better adapt to rare classes by effectively utilizing pre-trained knowledge.

Additionally, exploring newer transformer models like RoBERTa or GPT variants, which improve on BERT's architecture, or designing hybrid architectures combining the feature extraction capabilities of CNNs with the contextual understanding of transformers, could further enhance the accuracy and efficiency of intent classification systems. These innovations can lead to better handling of complex relationships within data, yielding improved predictions and overall performance.

Finally, investigating methods to adapt BERT for edge devices, such as through model quantization or distillation, would make high-performing models more accessible in resource-constrained environments. This adaptation would enable the deployment of robust intent classification systems in real-world applications, such as customer service bots or virtual assistants, without the need for extensive computational resources.

Ethics Statement

As with any AI-based system, intent classification models raise border societal and ethical considerations. The potential impact of this work includes:

Bais and Fairness:

- Bais in the dataset can propagate through the models, leading to unfair treatment of certain user groups or misinterpretation of inputs.
- Ensuring diverse and representative datasets is essential to mitigate this risk.

Privacy Concerns:

- The collection and use of user data to train these models raise privacy concerns.
- Implementing robust data anonymization and compliance with regulations like GDPR is necessary to protect users' rights.

Impact on Employment:

- As AI systems become integral to customer service and related industries, there may be implications for job displacement.
- Stakeholders must consider strategies to support workforce transition and upskilling.

Dependence on AI:

- Over-reliance on AI systems for decision-making can erode critical human judgment.
- Ensuring human oversight in sensitive applications is vital to maintain accountability.

References

M. Menda and G. S. Keerthi, "Intent Classification in Conversational System using Machine Learning Techniques," *Department of CSE, GVPCE*.

A. T. Al-Tuama and D. A. Nasrawi, "Intent Classification Using Machine Learning Algorithms and Augmented Data," *College of Computer Science and Information Technology, University of Kerbala, Iraq*.

A. Benayas, R. Hashempour, D. Rumble, S. Jameel, and R. C. de Amorim, "Unified Transformer Multi-Task Learning for Intent Classification With Entity Recognition," *Computer Science and Electrical Engineering Department, University of Essex, U.K.*, supported by Innovate U.K., under Grant 11422.

S. C. Han, S. Long, H. Li, H. Weld, and J. Poon, "Bi-directional Joint Neural Networks for Intent Classification and Slot Filling," *School of Computer Science, University of Sydney, Australia*.

H. Weld, X. Huang, S. Long, J. Poon, and S. C. Han, "A Survey of Joint Intent Detection and Slot-Filling Models in Natural Language Understanding," *The University of Sydney, Australia*.

B. Liu and I. Lane, "Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling," *Electrical and Computer Engineering, Carnegie Mellon University*.

E. Samunderu and M. Farrugia, "Predicting Customer Purpose of Travel in a Low-Cost Travel Environment—A Machine Learning Approach," *International School of Management (ISM), Dortmund, Germany; BCG Platinion Boston, USA*.