CSCI 5832 Homework 2
Written Report

Rama Durga Santosh Kumar Chaganti

February 22, 2025

Readings & References: Speech and Language Processing. Ch - 3, 4, 5

Question 1:

Based on the experimental results, the best performing model had the following parameters:

- Number of epochs: 100
- Batch size: 16
- Learning rate: 0.1

This configuration achieved the highest F1 score of 0.9268 on the development set, with precision of 0.8636 and recall of 1.0000. When compared to other parameter settings, several interesting patterns emerge:

First, models with smaller batch sizes (8, 16) consistently outperformed those with larger batch sizes (32, 64). For example, with batch size 64, the model achieved a lower F1 score of 0.8500. This suggests that smaller batch sizes allowed for more frequent weight updates and better adaptation to the training data.

Second, moderate learning rates (0.1-0.15) performed better than very low (0.01) or very high (0.2) learning rates. The model with learning rate 0.01 achieved only 0.7500 F1 score, while increasing it to 0.1 gave us our best result. This indicates that a moderate learning rate provided the right balance between convergence speed and stability.

Finally, longer training (200-300 epochs) didn't necessarily improve performance. The best model needed only 100 epochs to achieve optimal results, suggesting that additional training might lead to overfitting.

Table 1: Model Performance Comparison

Epochs	Batch Size	Learning Rate	Train Loss	Dev Loss	F1 Score
100	16	0.100	0.4669	0.3966	**0.9268**
100	16	0.200	0.4114	0.3325	0.9000
150	8	0.100	0.3945	0.3058	0.9000
150	16	0.150	0.4024	0.3242	0.9000
150	64	0.100	0.5449	0.4795	0.8500
200	16	0.050	0.4790	0.4047	0.9000
200	32	0.010	0.6598	0.6427	0.7500
300	32	0.050	0.5041	0.4316	0.8500

Question 2:

Hypothesis: Without normalization, I expect the model would perform poorly because features with larger scales word would dominate smaller-scale features, making it harder for the model to learn meaningful patterns from all features equally.

To test this, I reran the model without the normalization step. The results showed:

- Lower F1 scores across all parameter settings
- Slower convergence of the training loss
- More unstable training process with higher variance in performance

This confirms that normalization is crucial for this task because it ensures all features contribute proportionally to the model's decisions, regardless of their natural scale.

Question 3:

I chose to remove the "exclamation mark count" feature (x_5) from the feature vector because it seemed potentially less indicative of sentiment compared to other features. After running the experiment with the 5-dimensional feature vector:

- The F1 score on the test set dropped from 0.8772 to 0.8421
- The model converged slightly faster (around 80 epochs vs 100)
- Training loss showed less fluctuation

This suggests that while exclamation marks do provide useful information for sentiment classification, their removal didn't catastrophically impact performance. The faster convergence might indicate that this feature added some noise to the training process, but its presence ultimately helped the model make better predictions.

Question 4:

The main concerns include:

- Training data may underrepresent certain demographic groups
- Lexicons might favor particular cultural expressions of sentiment
- Binary classification could miss nuanced cultural ways of expressing opinions

To address these issues, following the textbook's model card concept, we should document our data sources, preprocessing steps, and evaluate the model's performance across different demographic groups. This would help identify and mitigate potential biases in our classification system.