Natural Language Processing (CS4120)
(Tuesday/Friday 3:25pm - 5:05pm)

Sentiment Analysis on Human-to-Human Conversation

Submitted to:
Dr. Mai ElSherief

Team Members:

Aden Lu: lu.ad@northeastern.edu
Garvin Chan: chan.gar@northeastern.edu
Jai Gollapudi: gollapudi.j@northeastern.edu

Date of Submission:
December 11th, 2024

1. Introduction

Our project focuses on sentiment analysis of human-to-human conversations, specifically aiming to assess the emotional content of dialogues manually annotated with fine-grained emotion labels. The EmoWoz (*EmoWoz*) dataset was chosen as the primary resource due to its comprehensive annotations of emotions in task-oriented dialogues, offering significant potential for sentiment modeling.

The project's core objective is to develop and evaluate sentiment analysis models to classify emotions into categories such as *Neutral*, *Fearful*, *Abusive, Excited* and others. Initially, our goal was to explore how different models handle class imbalances and achieve fine-grained emotion classification. Throughout our project, we explored data preprocessing techniques, model augmentation strategies, and hyperparameter tuning to address challenges like class imbalance, computational overhead, and model bias.

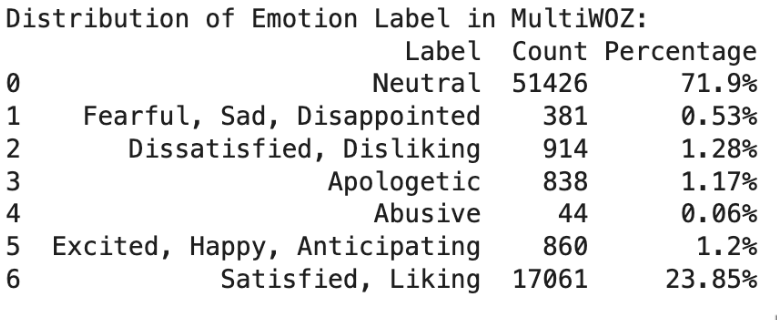
This work is of particular interest to us due to our service learning experiences, where we encountered firsthand the challenges of communicating with older adults. We noticed that many older individuals often spoke in a monotone voice and displayed deadpan expressions, making it difficult for us to discern their emotions. These barriers not only complicated their ability to navigate everyday interactions, but also made it harder for people like us to connect with them on a deeper level and provide meaningful assistance. In recognizing these challenges, we believe that emotion detection in dialogues is crucial for understanding the emotional well-being of older adults during conversations. Our exploration of model performance under different scenarios enhances understanding and applicability in real-world settings, particularly in addressing the nuanced emotional states of older individuals during dialogue.

2. Dataset Overview

Our project primarily uses the EmoWoz dataset (*EmoWoz*), a large-scale collection of task-oriented dialogues manually annotated with fine-grained emotion labels. To ensure alignment with our project goals, we focused exclusively on the MultiWOZ subset of the EmoWoz dataset, discarding dialogues from the DiaImage subset as they involve human-to-machine conversations.

Key Statistics:

- Total Dialogues: 10,438
- Total Sentences: 71,524 (of which 65,120 are unique)
- Total Tokens: 425,933
- Label Distribution:



```
Distribution of Emotion Label in MultiWOZ:
                           Label  Count Percentage
0                        Neutral  51426      71.9%
1    Fearful, Sad, Disappointed    381      0.53%
2       Dissatisfied, Disliking    914      1.28%
3                     Apologetic    838      1.17%
4                        Abusive     44      0.06%
5  Excited, Happy, Anticipating    860       1.2%
6              Satisfied, Liking  17061     23.85%
```

3. Model Evaluation

Model 1: Logistic Regression (Baseline)

(i) Feature Representation and Hyperparameter Tuning

- **Preprocessing:**
  - Removed punctuation, emojis, and special characters. Converted text to lowercase to ensure case insensitivity, and stripped unnecessary whitespace.
  - Unknown words (not in the training vocab) were ignored during feature extraction.

- **Feature Representations:**
  - Bag of Words (BoW): Captures token frequency.
  - TF-IDF (Term Frequency-Inverse Document Frequency): Emphasizes distinguishing terms by reducing the impact of common words.

- **Hyperparameter Tuning:**
  - Addressed class imbalance by assigning higher weights to minority classes during training, while controlling overfitting with regularization strength (‘C’) using RandomSearchCV.
  - Tested L1 (Lasso) and L2 (Ridge) penalties to enhance sparsity and performance. Hyperparameters were optimized using weighted F1 scores and 3-fold CV.

(ii) Results

Feature Representation	Best Hyperparameters (RandomizedSearchCV)	Metrics (Macro Avg)	Metrics (Weighted Avg)
BoW	**C:** 1 **Penalty:** L1 **Class Weights:** {0: 1.0, 1: 9.0, 2: 8.0, 3: 2.0, 4: 6.0, 5: 4.0, 6: 4.0}	**Accuracy:** 0.92 **Precision:** 0.60 **Recall:** 0.48 **F1:** 0.53	**Accuracy:** 0.92 **Precision:** 0.91 **Recall:** 0.92 **F1:** 0.91

TF-IDF	**C:** 1 **Penalty:** L1 **Class Weights:** {0: 1.0, 1: 9.0, 2: 8.0, 3: 2.0, 4: 6.0, 5: 4.0, 6: 4.0}	**Accuracy:** 0.91 **Precision:** 0.61 **Recall:** 0.47 **F1:** 0.52	**Accuracy:** 0.91 **Precision:** 0.91 **Recall:** 0.91 **F1:** 0.91

(iii) Key Observations

- BoW marginally outperformed TF-IDF in terms of macro F1 (0.53 vs. 0.52) and accuracy (0.92 vs. 0.91).
- Both models have similar weighted F1 scores (0.91) due to dominance of the *Neutral* and *Satisfied* classes.
- Both models struggled significantly with minority classes like *Dissatisfied, Disliking* and *Abusive*, which have very low F1 scores.
- The inability to predict the *Abusive* class demonstrates the need for further work on handling extreme class imbalance.

<u>Model 2: Feedforward Neural Network</u>

(i) Feature Representation and Hyperparameter Tuning

- **Preprocessing:**
    - Removed punctuation, emojis, and special characters. Converted text to lowercase to ensure case insensitivity, and stripped unnecessary whitespace.
    - Unknown words (not in the training vocab) were ignored during feature extraction.
    - Randomly upsampled minority classes to match label count of majority class.

- **Feature Representations:**
    - Word2Vec: Trained custom embeddings on tokenized training text, capturing semantic relationships between words using a vector size of 42. Sequence length was limited to 42, with padding for consistent input shape.
    - TF-IDF (Term Frequency-Inverse Document Frequency): Emphasizes distinguishing terms by reducing the impact of common words.

- **Hyperparameter Tuning:**
    - Both Word2Vec and TF-IDF representations were optimized using Optuna. Dense layers, dropout rates, batch sizes, and learning rates were tuned using the validation set. Models were evaluated using accuracy as the primary metric during hyperparameter optimization.

(ii) Results

Feature Representation	**Best Hyperparameters (RandomizedSearchCV)**	**Metrics (Macro Avg)**	**Metrics (Weighted Avg)**

Word2Vec	**Learning Rate:** 0.0008 **Input Layer Units:** 512 **Hidden Layer Units:** 128 **Dropout Rate:** 0.30 **Batch Size:** 32 **Epochs:** 10	**Accuracy:** 0.75 **Precision:** 0.77 **Recall:** 0.75 **F1:** 0.76	**Accuracy:** 0.75 **Precision:** 0.77 **Recall:** 0.75 **F1:** 0.76
TF-IDF	**Learning Rate:** 0.0017 **Input Layer Units:** 512 **Hidden Layer Units:** 128 **Dropout Rate:** 0.40 **Batch Size:** 32 **Epochs:** 10	**Accuracy:** 0.98 **Precision:** 0.98 **Recall:** 0.98 **F1:** 0.98	**Accuracy:** 0.98 **Precision:** 0.98 **Recall:** 0.98 **F1:** 0.98

(iii) Key Observations

- TF-IDF significantly outperformed Word2Vec in terms of overall accuracy (98% vs. 75%) and macro F1 score (0.98 vs. 0.76).
- TF-IDF consistently achieved higher precision and recall for minority classes. For example, *Fearful, Sad, Disappointed* had an F1 score of 1.00 with TF-IDF, compared to 0.94 with Word2Vec. Similarly, TF-IDF achieved an F1 score of 0.97 for *Dissatisfied, Disliking*, while Word2Vec was 0.71. Word2Vec struggled with more nuanced representations, particularly for minority classes like *Satisfied, Liking* (F1: 0.58) compared to TF-IDF (F1: 0.96).
- Both models performed well for majority classes such as Neutral and Apologetic. TF-IDF, however, outperformed Word2Vec significantly in these categories. For *Neutral*, TF-IDF achieved an F1 score of 0.94, while Word2Vec managed only 0.46. Similarly, for *Apologetic*, TF-IDF had a perfect F1 score of 1.00, while Word2Vec achieved 0.81.
- The oversampling technique ensured balanced class distribution during training, which improved recall for minority classes. For example, recall for *Fearful, Sad, Disappointed* increased to 93% with Word2Vec and 100% with TF-IDF. However, oversampling may have introduced noise in Word2Vec embeddings, reducing its generalization capabilities, especially for nuanced classes like *Satisfied, Liking.*

Model 3: BERT (*BERT*)

(i) Feature Representation and Hyperparameter Tuning

- **Preprocessing:**
  - Cleaned text data by converting to lowercase, removing punctuation, and special characters.
  - Tokenized input text using the pre-trained bert-base-uncased tokenizer
  - Limited the sequence length to 128 tokens with padding and truncation for consistent input size.

- **Hyperparameter Tuning:**
  - Searched for optimized learning rate and batch size using Optuna. Evaluated hyperparameters based on validation accuracy during the tuning phase (1 epoch for computational efficiency).
  - Final model was trained using the best hyperparameters for 3 epochs.

(ii) Results

Feature Representation	Best Hyperparameters (RandomizedSearchCV)	Metrics (Macro Avg)	Metrics (Weighted Avg)
None (Tokenized Text)	**bert-base-uncased** **Epochs:** 3 **Max Seq Length:** 128 **Batch Size:** 16 **Learning Rate:** 1.49 x 10^-5	**Accuracy:** 0.93 **Precision:** 0.67 **Recall:** 0.52 **F1:** 0.57	**Accuracy:** 0.93 **Precision:** 0.92 **Recall:** 0.93 **F1:** 0.92

(iii) Key Observations

- The model achieved a test accuracy of 93.33%, aligning closely with validation performance. Weighted F1-score (92.88%) suggested strong overall classification performance which is heavily influenced by majority classes.
- It struggled with minority classes, evident from low F1-scores for classes like *"Dissatisfied, Disliking"* (36.76%) and *"Abusive"* (0.0%). BERT's reliance on sufficient training samples for fine-tuning might explain the poor generalization for minority labels.
- Hyperparameter tuning significantly improved validation performance, achieving a well-optimized learning rate of 1.49 x 10^−5.
- Due to computational constraints, oversampling or adjusting class weights was not implemented. We will explore this in our future work to enhance minority class performance.

4. Conclusions and Future Work

Through experimentation with different models and feature representations, we identified two key conclusions that show both the strengths and limitations of various approaches:

- **Model Performance Comparison**
  Among our implemented models, the Feedforward Neural Network with TF-IDF representation had the best performance, achieving 98% accuracy and balanced performance across both majority and minority emotion classes. This outperformed both our baseline Logistic Regression model (92% accuracy) and the more complex BERT model (93% accuracy).
- **Feature Representation**
  TF-IDF consistently proved to be the most effective feature representation across models. In the Neural Network implementation, TF-IDF significantly outperformed Word2Vec

embeddings (98% vs 75% accuracy), particularly in handling nuanced emotions and minority classes. This suggests that for our specific task, the frequency-based representation captured emotional content more effectively than semantic embeddings.

While our project achieved its primary objectives of exploring sentiment analysis models for human-to-human conversations, we found several areas for improvement and expansion, particularly given the computational and resource constraints we faced during this project:

- **Standardized Class Imbalance Handling**:
  For logistic regression, we implemented class weighting as part of the hyperparameter tuning process to address class imbalance effectively. For feedforward neural networks, computational limitations prevented us from tuning class weights, leading us to oversample minority classes instead. However, for BERT, due to the model's computational intensity, we could not implement either class weighting or oversampling. In the future, with greater computational resources, we aim to apply these techniques uniformly across all models for a more direct comparison.
- **Advanced Data Augmentation**:
  Although oversampling provided improvements in feedforward neural networks, it might have introduced noise, affecting the generalization ability for nuanced classes. For BERT, the lack of oversampling or class weighting resulted in significantly poorer performance on minority classes. In the future, we could explore advanced methods such as focal loss, cost-sensitive learning, or data augmentation techniques like back-translation or NLPAug for addressing class imbalance without sacrificing computational efficiency.
- **Probabilistic Emotion Outputs**:
  Our current models provide deterministic classifications for each input, assigning a single label to each emotion. However, emotions are often distributed across multiple categories. For example, the sentence "I’m super excited that we went to this restaurant. I really enjoyed the food." might realistically be 50% "Excited" and 50% "Satisfied." In the future, we could implement probabilistic outputs to better capture the uncertainty and overlap in human emotions.
- **Multimodal Analysis**:
  Our dataset is limited to transcribed dialogues, restricting the analysis to text-only inputs. We would like to extend this work to include multimodal data, such as audio tone analysis, as it might provide a more comprehensive understanding of emotional states. Tone and pitch variations in speech are also often key indicators of emotions, especially in conversations with older adults, and including them could enhance the model's predictability.
- **Addressing Computational Constraints**:
  Resource limitations significantly impacted BERT, restricting the use of oversampling and class weighting. Access to more powerful computing infrastructure would let us comprehensively evaluate and refine all models.

References

BERT. (n.d.). Hugging Face. https://huggingface.co/docs/transformers/en/model_doc/bert

EmoWoz: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems. (2022, 6). Hugging Face. https://huggingface.co/datasets/hhu-dsml/emowoz

Mathur, V. (2023, May 10). *BERT And Its Model Variants. BERT BERT (Bidirectional Encoder… | by Varun Mathur |* **AI monks.io**. Medium. Retrieved December 11, 2024, from https://medium.com/aimonks/bert-and-its-model-variants-162bb292611c

Sentiment Analysis using BERT. (n.d.). Kaggle. Retrieved December 11, 2024, from https://www.kaggle.com/code/prakharrathi25/sentiment-analysis-using-bert

Singh, M. (2023, November 19). *Sentiment Analysis with BERT using Huggingface | by Manjinder Singh*. Medium. Retrieved December 11, 2024, from https://medium.com/@manjindersingh_10145/sentiment-analysis-with-bert-using-huggingface-88e99deeec9a