Utah Wellbeing Project Open Comments Text Classification

Nicolas Holden
Department of Mathematics and Statistics
Utah State University (USU)
Logan, USA
nicolas.holden@usu.edu

***Abstract*—The Utah Wellbeing Project assesses and monitors wellbeing across Utah communities to support informed city planning and decision-making. In 2024, over 15,000 respondents from 49 Utah cities participated in an online survey. The survey included open-ended questions, such as what residents value most about their city and what improvements they would like to see. This effort resulted in 26,414 comments, which were shared with city leaders upon survey completion. However, organizing these comments for analysis and actionable insights is a time-intensive task. To address this challenge, this project applies natural language processing (NLP) techniques and sequence-based neural networks to classify comments into approximately 24 dominant topics, such as transportation, safety, recreation, and healthcare. Results indicate that gated recurrent units (GRUs) combined with pre-trained Word2Vec embeddings achieve modest performance in comment classification, providing a foundation for more efficient and insightful analysis of community feedback.**

***Index Terms*—GRU, classification, supervised, natural language processing**

I. Introduction and Goals

A. Motivation

The Utah Wellbeing Project, led by Dr. Courtney Flint, has been tracking wellbeing across Utah communities and cities since 2019, providing valuable insights for researchers and city leaders. In 2024 alone, the project collected 15,553 comments, resulting in 26,414 text-based answers. Since 2020, over 38,000 comments have been recorded statewide. These comments address a wide variety of topics, including transportation, safety, housing, government administration, healthcare, and more. Ensuring that each respondent's voice is heard and considered is essential for fostering wellbeing improvements across communities.

A primary objective of the survey is to organize these open-ended comments into meaningful topics. This organization allows researchers and city leaders to efficiently analyze patterns and identify common themes, supporting data-driven decision-making. However, categorizing comments by topic presents significant challenges, as many responses span multiple themes. For instance, comments about housing often intersect with concerns about growth and development, making it difficult to assign a single category. Nevertheless, classifying each comment by its 'dominant' topic facilitates easier processing and interpretation of the data.

Logical organization of comments is critical for actionable insights at both the state and local levels. For city leaders, it enables quick identification of relevant feedback without requiring exhaustive searches, thereby helping them understand residents' priorities and concerns more efficiently. For researchers, the data offers a foundation to study broader trends across cities and explore how wellbeing varies by topic throughout Utah. This capability ultimately helps all invested parties address the needs of communities more effectively.

B. Goals

To accomplish the organization of comments, this project uses sequence-based neural networks to classify and organize comments from the 2024 survey into one of 24 topics. Each comment has already been manually labeled with a dominant topic by human researchers, thus enabling supervised classification. The primary goal of the project is to achieve the best classification metrics, generally defined by the macro F1 score. Secondary objectives of the project aim to compare the performance of difference sequence based models and the effects of fine tuning. Additionally, investigating the data and how topics overlap is insightful for fine-tuning labeled topics.

II. Related Work

The application of deep learning techniques for text classification, particularly sequence based models, has gained significant attention in recent years. One such architecture is the Gated Recurrent Unit (GRU) network. Cho et al. (2014) introduced GRUs as an alternative to Long Short-Term Memory (LSTM) networks for sequence modeling, noting their efficiency and ability to mitigate the vanishing gradient issue [1]. Additionally, convolutional neural networks (CNNs) have been used for sentence classification, with Kim (2014) demonstrating the efficacy of CNNs for text classification tasks [2].

For natural language processing (NLP) tasks, word embeddings such as word2vec have greatly increased performance in models. Mikolov et al. (2013) proposed word2vec as a method to learn dense vector representations for words, capturing semantic meanings of words [3]. This approach has shown in many models at improving the model accuracy.

III. Methods

The analysis was conducted using TensorFlow and Keras Python libraries for developing and evaluating deep neural

network models. A Naïve Bayes classifier, implemented via scikit-learn, served as the baseline model.

The analysis process followed this general flow:

1) Data Cleaning: Grouping classes and collecting data.
2) Data Exploration: Understanding class distributions and imbalance and train/test splitting.
3) Naïve Bayes Baseline: Establishing a baseline performance using a simple classification model.
4) Comparison of sequence-based neural networks: Evaluating different architectures including LSTM, RNN, and GRUs.
5) Comparison of different GRU model architectures: Testing various GRU configurations to optimize performance.
6) Fine-tuning a GRU model: Hyper-parameter tuning to maximize model performance and generalization.
7) Evaluation of model performance: Evaluate model performance through various metrics and visual diagnostics.

A. Evaluation Metrics

Given the large number of classes (24) and significant class imbalance, the macro F1 score was the primary evaluation metric for comparing model performance. Additionally, model loss over epochs was monitored in each step of the process to identify over fitting.

B. Data Splitting and Reproducibility

The dataset was split into training and testing subsets (80/20 split) using a stratified split for consistent comparison across models. During hyper parameter tuning, the training data was further divided into training and validation subsets (80/20 split). All evaluation metrics reported reflect performance on the test set. To ensure reproducibility, random seeds were set in TensorFlow prior to each model run.

C. Word Embeddings

For the Naïve Bayes classifier, comments were converted into numerical features using the CountVectorizer() from scikit-learn. For sequence-based neural networks, pretrained Word2Vec embeddings (Google News, 300 dimensions) were applied using the Gensim library.

D. Fine-Tuning Process

Model fine-tuning was performed using the keras_tuner package, focusing on optimizing the following hyper parameters:

- Learning Rate: Ranging from 0.1 to 0.0001.
- Beta 1 (Adam Optimizer): Ranging from 0.9 to 0.999.
- Number of Hidden Units (GRU): Ranging from 32 to 512.
- Dropout Rate: Ranging from 0.0 to 0.5.

A random search of 20 models was conducted, with the average performance over three runs per model recorded for comparison. Early stopping was applied with a maximum of 50 epochs to ensure efficient training and prevent over fitting.

IV. Experiments

The code for this project can be found on GitHub [1]

A. Dataset Description

The dataset used for this analysis originates from the Utah Wellbeing Project and was obtained with permission from Dr. Courtney Flint. Due to the protected nature of human-subject research data, the dataset cannot be shared publicly. Only non-identifying, generic comments may be shared on a limited basis.

The dataset initially contained 26,414 comments across 24+ labeled topics. During data cleaning prior to this project, the topics were refined to 24 specific categories by correcting spelling errors during data entry. Additionally, 85 blank comments were removed, resulting in a final dataset of 26,329 comments with annotated dominant topics.

The dataset exhibits significant class imbalances. For example, Transportation is the most frequently occurring topic with 3,044 comments, while Substance Misuse is the least represented, with only 117 instances. The distribution of the number of words per comment appears approximately exponential, with the mode falling between 5 and 10 words. A small subset (1.3%) of comments contains 100 words or more.

Prevalence of topics also varied by the original survey question. Three questions were asked in the survey to which these responses were recorded from:

1) What do you value most about living in [your city]?
2) Is there anything that could be done to improve wellbeing in [your city]?
3) Is there anything else you'd like to tell us about wellbeing in [your city]?

Respondents most frequently discussed City/Town Character, Social Climate, Location, Peace and Quiet, and Nature and Beauty topics when responding to the first question. Respondents most frequently discussed Transportation, Environment and Resources, Government, Housing, Recreation and Tourism, and Retail and Business topics when responding to the second question. Lastly, the third question experienced a higher mixture of topics with many respondents indicating nothing or that they were unsure.

B. Baseline Methods

To establish a rough baseline using a simple machine learning model, the textual comments were transformed into numerical features using the CountVectorizer() from the scikit-learn library. Following this, a Naïve Bayes classifier was applied to the transformed data, both with and without up-sampling the minority classes. The performance metrics of the baseline models are presented in Table I. While up sampling does not help the precision of the model, it does significantly improve the recall.

[1] https://github.com/Needle13/cs5640_final/tree/main

TABLE I
BASELINE PERFORMANCE METRICS WITH A NAÏVE BAYES CLASSIFIER

Metric	Without Up Sampling	With Up Sampling
Macro Precision	0.63	0.60
Macro Recall	0.39	0.55
Macro F1	0.42	0.56
Overall Accuracy	0.55	0.60

C. Sequence Models

To evaluate the effectiveness of sequence-based models, SimpleRNN, LSTM, and GRU models from Keras were initially trained and compared based on their macro F1 scores. Each model was further modified with three different architectures: a Dropout Layer, a Bidirectional Layer, and a CNN Layer. The resulting macro F1 scores for each configuration are presented in Table II.

The GRU model consistently outperformed both the SimpleRNN and LSTM models across all architecture modifications. Among all configurations, the GRU model with Dropout achieved the highest macro F1 score of 0.72, a large improvement over the Naïve Bayes classifier.

TABLE II
MACRO F1 SCORES WITH SEQUENCE BASES NEURAL NETWORKS

Macro F1 Scores	LSTM	RNN	GRU
Base Model	0.68	0.48	0.69
With Dropout	0.69	0.36	0.72
Bidirectional	0.68	0.37	0.69
With CNN	0.66	0.37	0.68

D. GRU Models

Further experimentation with the GRU architecture involved exploring additional layer combinations to capture more complexity in the data. Despite these modifications, the performance did not improve significantly. The results of the various model configurations are presented in Table III. Among the models tested, the basic GRU and GRU with Dropout achieved the best performance.

Increasing model complexity by adding extra layers, incorporating CNNs, or using a Bidirectional GRU did not result in any substantial improvements. Although the performance of the base GRU and the GRU with Dropout models was nearly identical, analysis of the model loss over epochs revealed that the base GRU model was prone to overfitting after approximately 15 epochs. Based on this observation, the GRU with Dropout model was selected as the final architecture.

E. Comment Length Tuning

An additional experiment was conducted to investigate how varying the maximum comment length would affect the model's ability to correctly classify topics. The hypothesis was that truncating the comments might remove extraneous information, with the assumption that the dominant topic would typically be established early in the comment.

TABLE III
PERFORMANCE METRICS FOR DIFFERENT MODIFICATIONS OF GRU MODEL

Modifications	Accuracy	Precision	Recall	F1
Base GRU	0.72	0.72	0.69	0.70
With Dropout	0.72	0.72	0.69	0.70
Bidirectional	0.71	0.70	0.68	0.69
CNN, Dropout, Bidirectional	0.72	0.71	0.67	0.68
CNN, Dropout	0.71	0.70	0.67	0.68
CNN	0.71	0.69	0.66	0.67
CNN (2x), Dropout (2x)	0.69	0.67	0.62	0.63

However, the results contradicted this hypothesis. Including the full comment length generally led to better performance compared to truncating the comment, as shown in Figure 1. This suggests that the additional context provided by longer comments is valuable for accurately identifying the dominant topic.

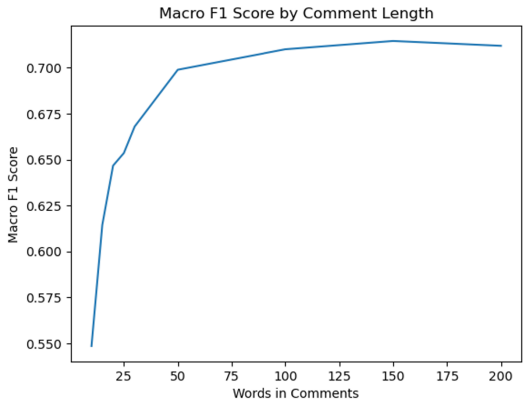

Fig. 1. Effect of Max Comment Length on Macro F1 Scores

F. Fine-tuning

Using the keras_tuner package and RandomSearch, the best model configuration was identified, which included a learning rate of 0.001, a Beta 1 value of 0.999, 224 hidden units, and a dropout rate of 0.2. The performance metrics of this fine-tuned model, compared to both the Naïve Bayes and untuned GRU models, are presented in Table IV.

The results show that the neural network models significantly outperform the Naïve Bayes classifier. Additionally, the fine-tuned model demonstrates a slight improvement over the untuned GRU model.

G. Evaluation

The fine-tuned GRU model was then evaluated in detail using the test dataset. The training and testing model loss curves are shown in Figure 2. Although the Dropout layer helps mitigate overfitting, there is a slight decline in test set model loss as the number of epochs exceeds 30.

TABLE IV
PERFORMANCE COMPARISON OF NAÏVE BAYES, UNTUNED GRU, AND FINETUNED GRU MODELS

Metric	Naïve Bayes	Untuned GRU	Finetuned GRU
Macro Precision	0.60	0.72	0.73
Macro Recall	0.55	0.72	0.70
Macro F1	0.56	0.69	0.72
Overall Accuracy	0.60	0.70	0.74

By comparing both the validation loss and accuracy, it was determined that 30 epochs provided the optimal balance, yielding the highest model accuracy. This number of epochs was chosen as the final stopping point for maximum model generalization.

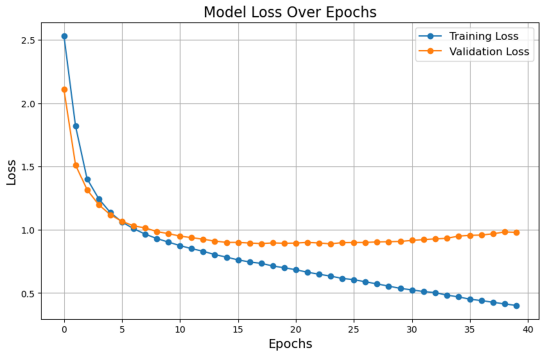

Fig. 2. Training and Testing Model Loss over Epochs

A deeper investigation into the F1 scores and confusion matrix of the model reveals both its strengths and areas for improvement. The accuracy metrics for each class are provided in Table V.

From the results, it is evident that the Transportation class is the easiest to classify, while the Other category proves to be the most challenging. This analysis highlights the model's ability to correctly classify certain topics while struggling with more ambiguous or less distinct categories.

A confusion matrix was also analyzed on the test set as seen in Figure 3. The matrix revealed that comments related to Growth and Development are sometimes misclassified as Housing comments, and vice versa. Similarly, City/Town Character comments are occasionally misclassified as Social Climate, while Social Climate comments are sometimes misclassified as Government.

Upon reviewing these misclassified comments, it becomes clear why the model struggles to differentiate between these categories. There is significant overlap between these classes, with comments that could reasonably fit into either category. While the model does occasionally misclassify comments entirely, the majority of misclassifications occur within these "gray" areas, where comments could belong to multiple topics.

V. CONCLUSION

Among the sequence-based neural networks evaluated, the GRU with Dropout achieved the highest macro F1 scores.

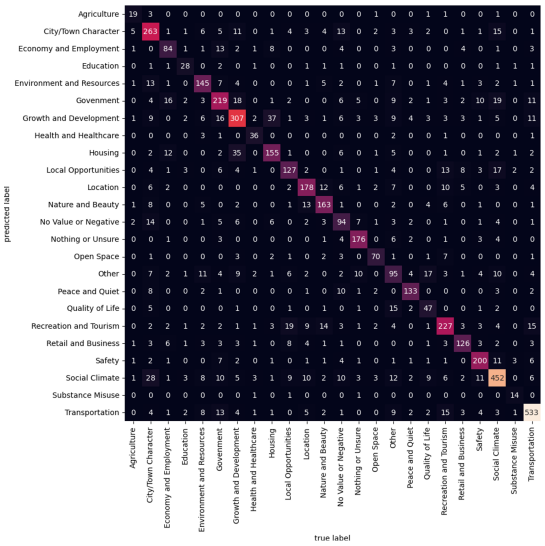

Fig. 3. Confusion Matrix on Test Set

More complex GRU architectures did not yield improvements in performance. Fine-tuning the model's hyperparameters provided a slight performance boost, while shortening the comment length actually reduced the model's effectiveness. The Dropout layer proved effective in mitigating overfitting. Overall, the final model achieved an improvement of approximately 14% in accuracy and 16% in macro F1 score compared to the Naive Bayes classifier.

While the overall accuracy was 0.74, model performance varied significantly across classes. A notable challenge arose when comments referenced multiple topics, as they could reasonably be classified into more than one category.

Future work on this project could explore several avenues for improvement, including:

- Attention-based models to better capture context and relationships in the comments.
- Investigating the impact of different pre-trained word embeddings.
- Experimenting with up sampling techniques to address class imbalances and improve model performance.

Additionally, while this project focused on assigning a single topic to each comment, the Utah Wellbeing Project could benefit from several extensions:

- Assigning multiple topics to a single comment, enabling more granular searches of the data.
- Utilizing text localization techniques to detect and extract specific themes or topics within comments.
- Utilizing Large Language Models (LLMs) to generate topic-specific summaries for individual communities, cities, or regions.

Given the large volume of data collected, organizing and analyzing these comments can greatly enhance the ability of local and state leaders to address relevant issues. Natural Language Processing and Sequence-based Neural Networks offer a promising framework for efficiently classifying and

TABLE V
PERFORMANCE METRICS FOR ALL CLASSES

Class	Precision	Recall	F1	Support
Transportation	0.87	0.88	0.87	609
Nothing or Unsure	0.88	0.84	0.86	210
Peace and Quiet	0.82	0.85	0.83	157
Safety	0.82	0.80	0.81	250
Social Climate	0.76	0.81	0.78	561
Nature and Beauty	0.78	0.77	0.77	213
Open Space	0.76	0.79	0.77	89
Health and Healthcare	0.82	0.72	0.77	50
Retail and Business	0.73	0.79	0.76	159
Location	0.75	0.76	0.75	235
Recreation and Tourism	0.71	0.74	0.73	306
Growth and Development	0.70	0.74	0.72	417
City/Town Character	0.76	0.68	0.72	387
Environment and Resources	0.73	0.71	0.72	204
Substance Misuse	0.82	0.61	0.70	23
Housing	0.69	0.71	0.70	217
Government	0.66	0.69	0.67	319
Local Opportunities	0.64	0.69	0.66	183
Education	0.70	0.62	0.66	45
Economy and Employment	0.67	0.64	0.65	132
Agriculture	0.73	0.58	0.64	33
No Value or Negative	0.61	0.53	0.57	178
Quality of Life	0.62	0.52	0.56	91
Other	0.49	0.48	0.48	198

understanding community feedback, with significant potential for further improvements.

The code for this project can be found on GitHub.

REFERENCES

[1] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." Available: https://arxiv.org/pdf/1406.1078v3

[2] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *arXiv.org*, 2014. https://arxiv.org/abs/1408.5882

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Neural Information Processing Systems*, 2013. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html