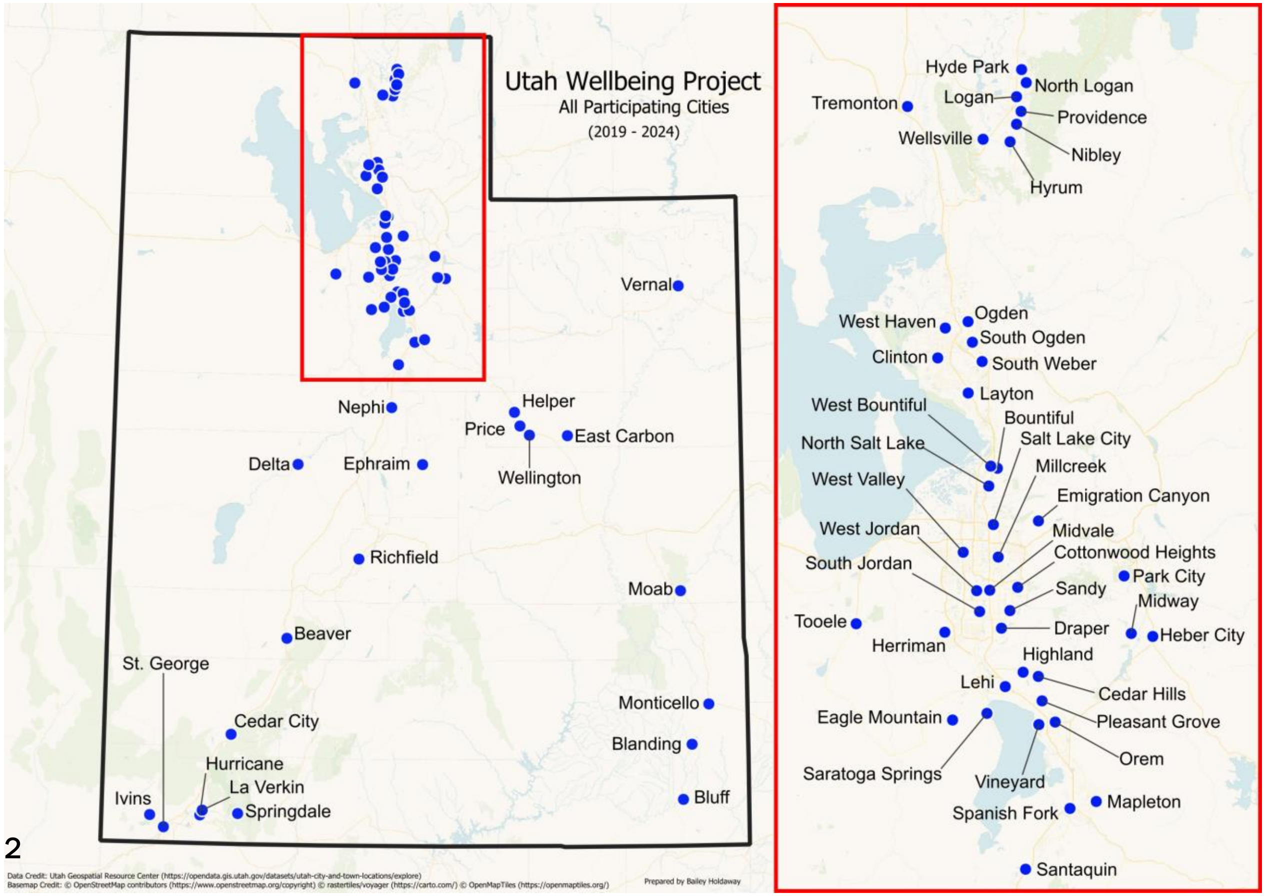
Utah Wellbeing Project Open Comments Text Classification

Nicolas Holden

Utah Wellbeing Project



- 49 Participating Cities in 2024
- 15,533 Survey Respondents in 2024
- 3 Primary Open Ended-Comments
- Aims to improve and track wellbeing across Utah communities

Open Comment Questions

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

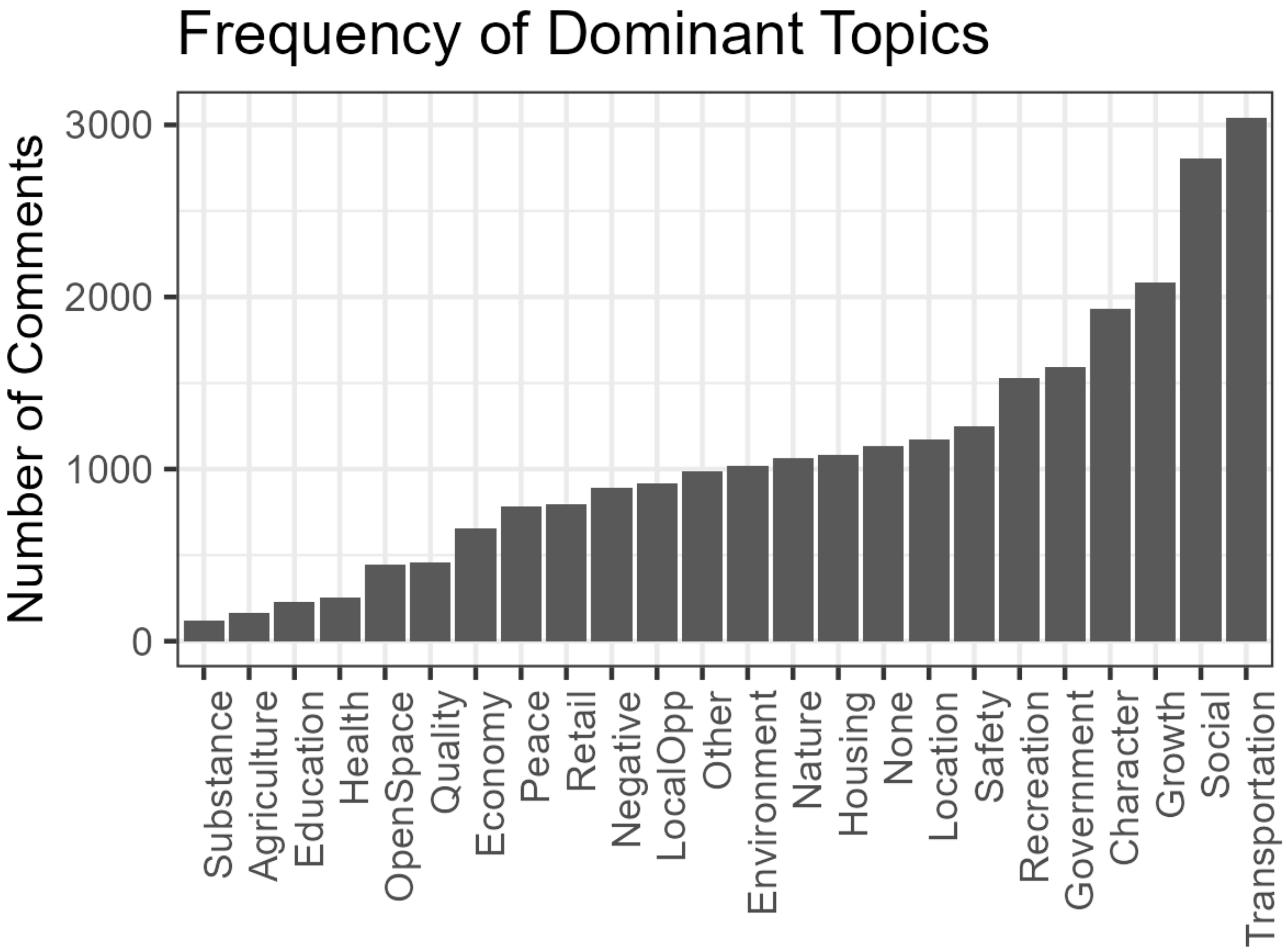
Sample Comments

Goal: Organize comments into dominant categories to give back to city leaders. (supervised classification)

Comment	Dominant Category
I love how people love living in [their city]. Most people seem to take care of their yards, homes, families	Quality
Police accountability and keeping the cost of living low.	Government
Nice neighborhoods, well-managed city, low crime rate	Social
Trail systems	Recreation

Pre-Labeled Dominant Topics

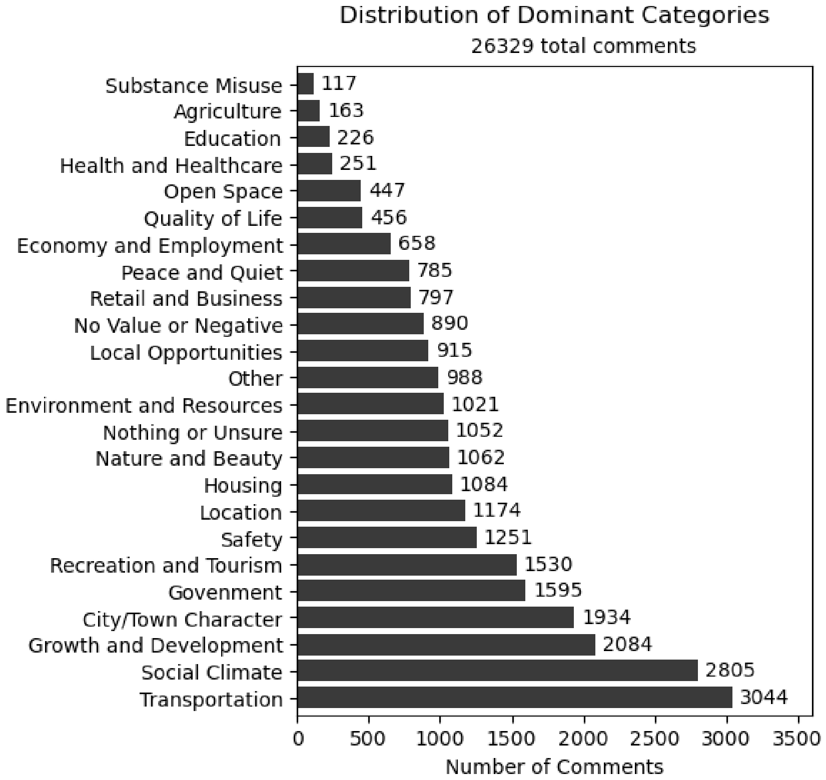
- 26,414 comments already labeled by a team of human researchers
- 24(ish) topics
- Unbalanced data classes

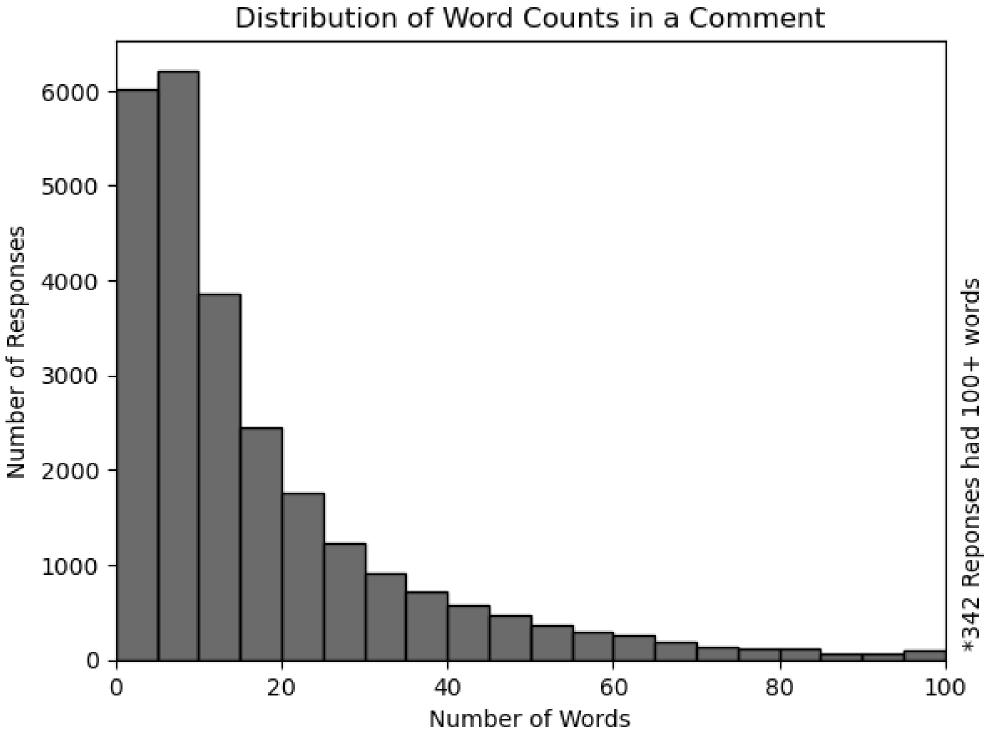


Methodology

1. Data cleaning
2. Data exploration
3. Naïve Bayes Baseline
4. Sequence Model Architecture Comparisons
5. GRU Model Architecture Comparisons
6. Fine-tuning
7. Evaluation and Assessment

Data Exploration





Naïve Bayes Baseline

- Utilized
  - CountVectorizer() from scikit-learn to embed features
  - SMOTE (up sampling) from imblearn to improve class imbalance

	Unbalanced Classes	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

Sequence Model Comparison

Ran with
- Word2vec pretrained embeddings from Google News 300 via genism
- 30 epochs, 256 batch size, default Adam optimizer from keras
- Hidden Layer size of 100
- Dropout of 0.2
- CNN with 32 filters, kernel size of 3, same padding
- Max comment length of 100

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

GRU Models generally performed the best

30 Epochs was induced overfitting. 10-15 epochs produced better generalizability

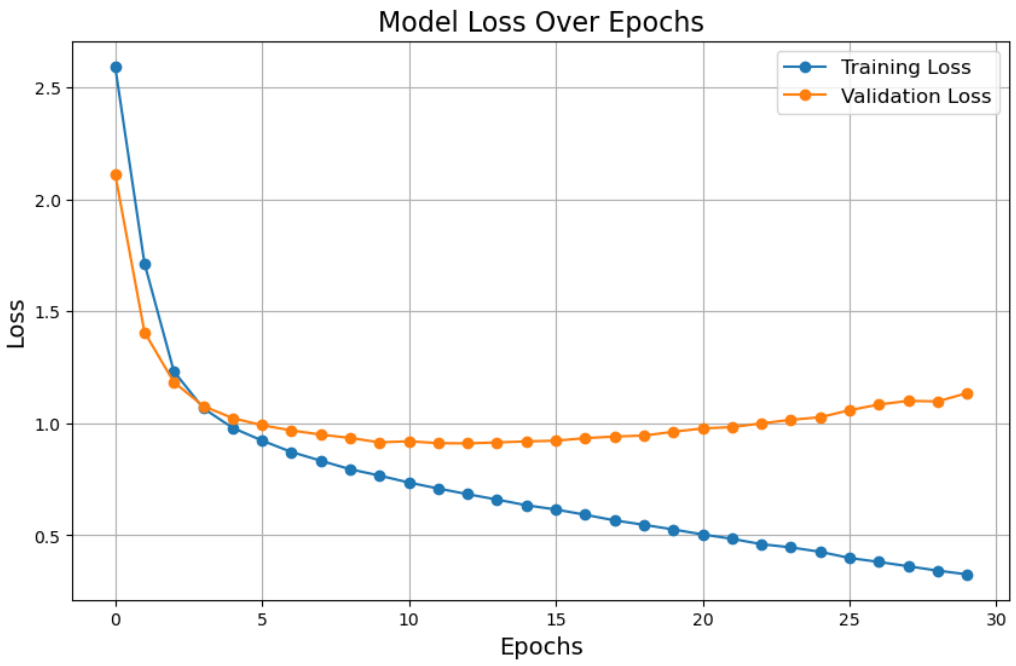
GRU Model Comparison

While the base GRU and w/ Dropout GRU model performed equally, the dropout model was much more resistant to overfitting.

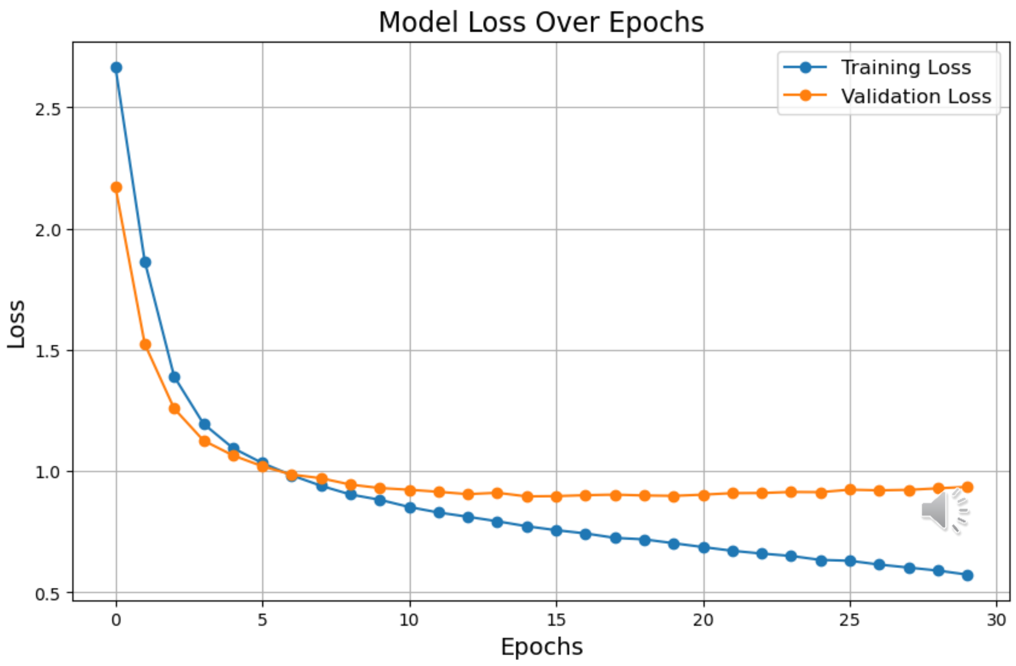
Modifications	Accuracy	Precision	Recall	F1
Base GRU	0.72	0.72	0.69	0.70
With Dropout	0.72	0.72	0.69	0.70
Bidirectional	0.71	0.70	0.68	0.69
CNN & Dropout & Bidirectional	0.72	0.71	0.67	0.68
CNN & Dropout	0.71	0.70	0.67	0.68
CNN	0.71	0.69	0.66	0.67
CNN (2x) & Dropout (2x)	0.69	0.67	0.62	0.63

Dropout Effects

GRU Models without Dropout were prone to overfitting



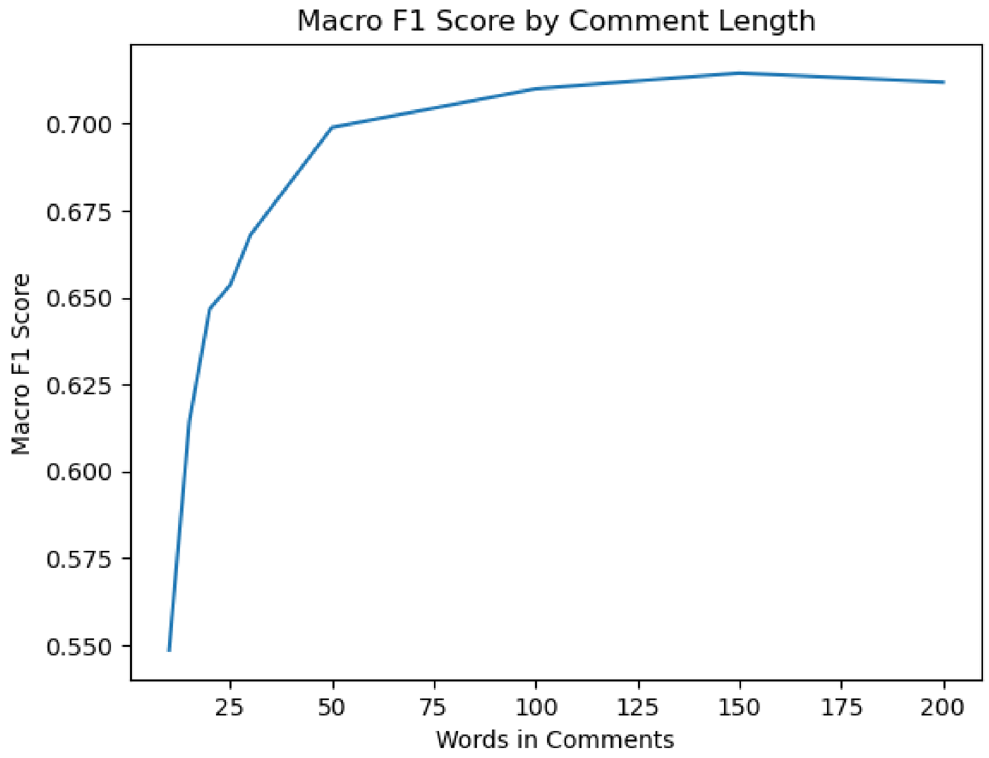
GRU Models with Dropout were more generalizable and had less variance



Comment Length Effects

Cutting the comments too short removed necessary context for the comments.

However, after ~150 words, more context does not help the classification.



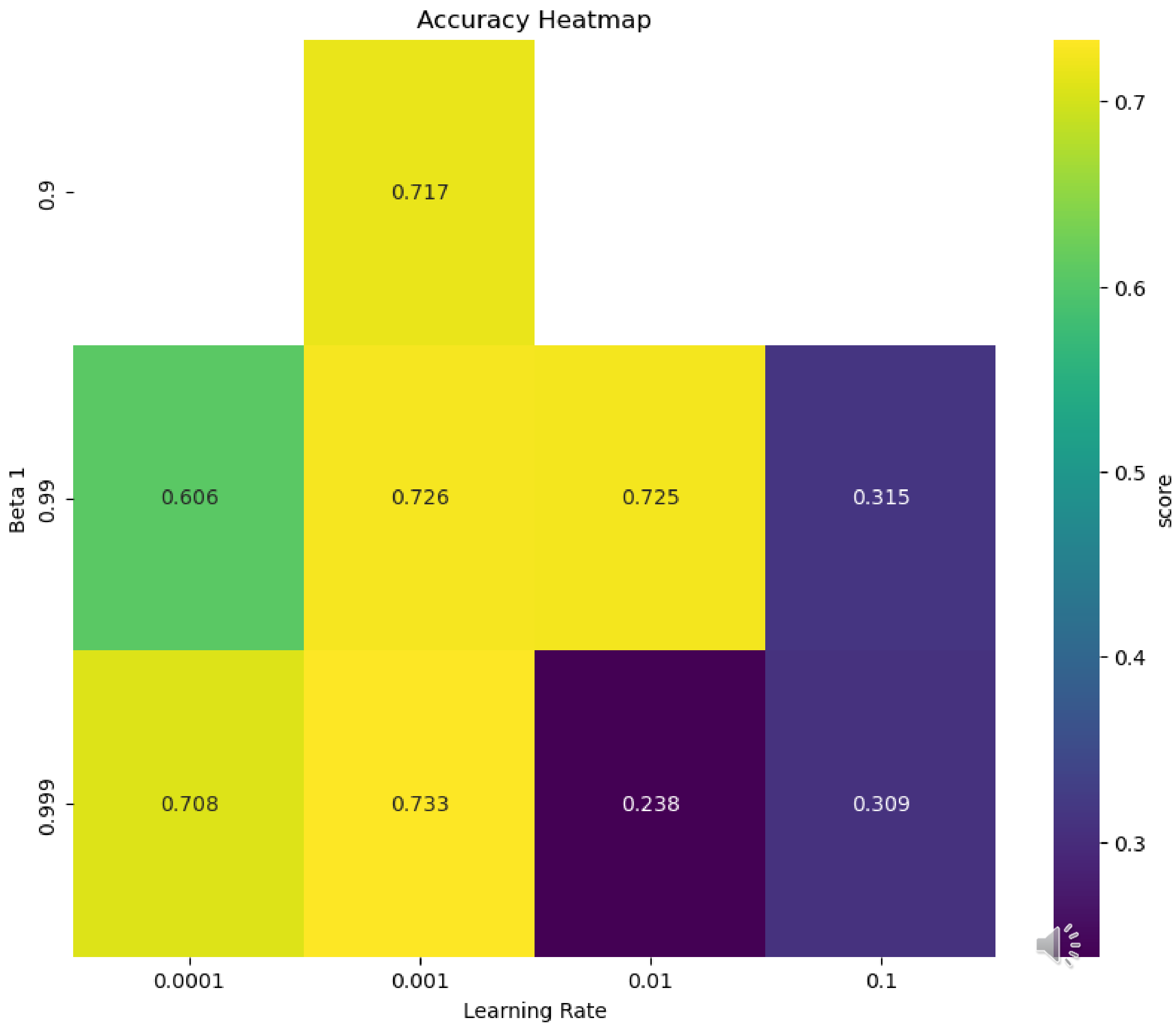
Finetuning

Finetuned

- Learning rate: [0.1, 0.01, 0.001, 0.0001]
- Beta 1: [0.9, .99, 0.999]
- Hidden Layer Units: [32 -> 512 units]
- Dropout Rate: [0, 0.1, 0.2, 0.3, 0.4, 0.5]
- Using 3-fold validation
- 20/900 models tested

Best Model:

- Learning rate: 0.001
- Beta 1: 0.999
- Hidden Layer units: 224
- Dropout Rate: 0.2



Evaluation

F1 Scores:

- Min: Other – 0.48
- Max: Transportation – 0.87

Frequent Misclassifications (Actual -> Predicted)

- Growth and Development <-> Housing
- City/Town Character -> Social Climate
- Social Climate -> Government
- Growth and Development -> Government

	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

Conclusions & Future Work

Conclusions

- GRU significantly outperformed a Naïve Bayes approach.
- GRU model outperformed other sequence based neural networks.
- Finetuning the GRU slightly improved results. Dropout mitigated overfitting.
- Model misclassifications generally coincide with unclear topic boundaries/comment could often belong in multiple categories.

Possible Modification on existing work:

- Use of Attention Based Models
- Effect of different Word embeddings (GloVe)
- Effect of up sampling for deep learning architectures

Possible Future Projects

- Detecting multiple topics/one at a time.
- Text localization and extraction of topics within a sentence.
- Automatic topic summarization of comments.

Thanks

References

- K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." Available: https://arxiv.org/pdf/1406.1078v3
- Y. Kim, "Convolutional Neural Networks for Sentence Classification," *arXiv.org*, 2014. https://arxiv.org/abs/1408.5882
- 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/9aa42b31882ec03 9965f3c4923ce901b-Abstract.html

Special thanks to Dr. Courtney Flint and all those researchers at the Utah Wellbeing Project for collecting, annotating, and providing the data set and for the work they do.

Utah WELLBEING PROJECT