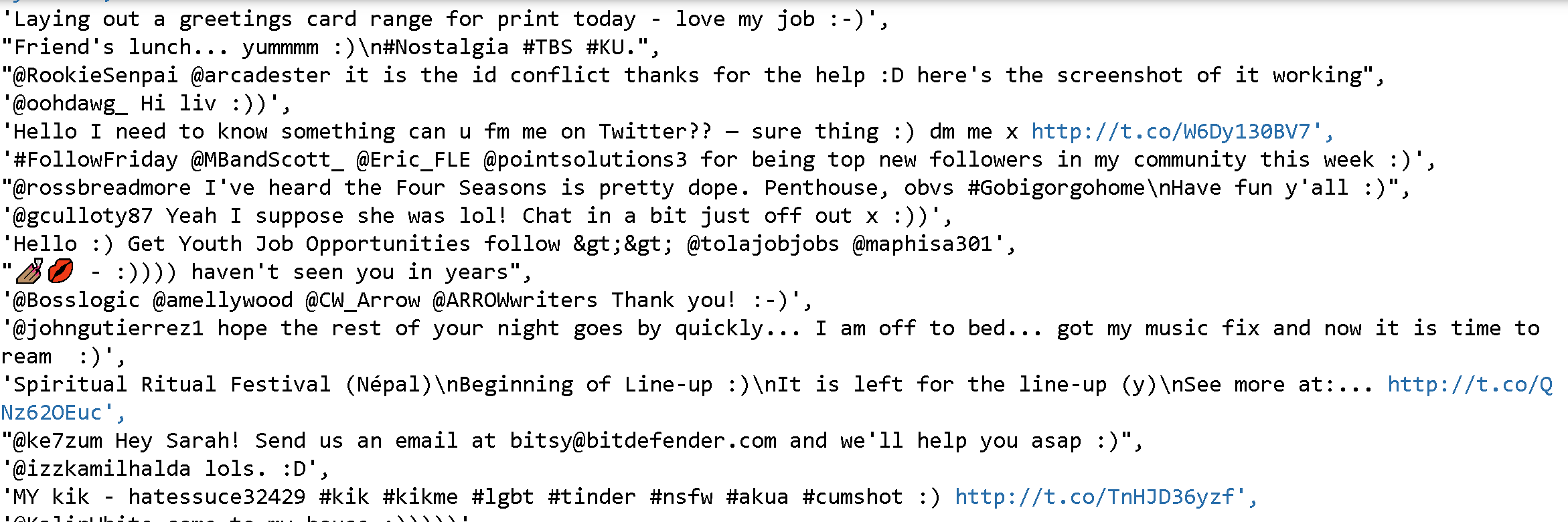
**Explainability of models in natural language processing for sentiment analysis**

**Purpose-**

Building explainable systems is a critical problem in the field of Natural Language Processing (NLP) since most models provide little to no explanations for their predictions. In most of the cases, the fine-grained information is often ignored, and the models do not explicitly generate the human-readable explanation. Applying modern NLP for real-world applications demands interpretability and to make the system more robust. This project aims to use established explainability tools such as lime and captum in various sentiment analysis tasks to show what factors lead to the model’s prediction. Also, zero in on the false positive and false negative results to see what specific words are causing the models to classify them as wrong and finally compare the models making the predictions.

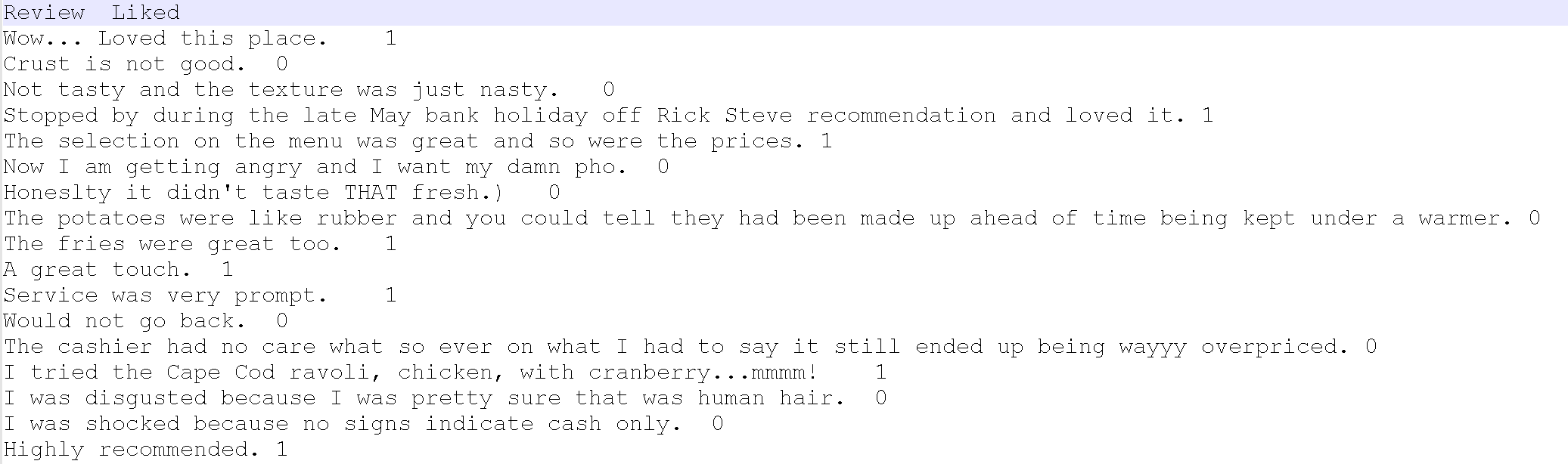
**Datasets-**

Positive and negative tweets:



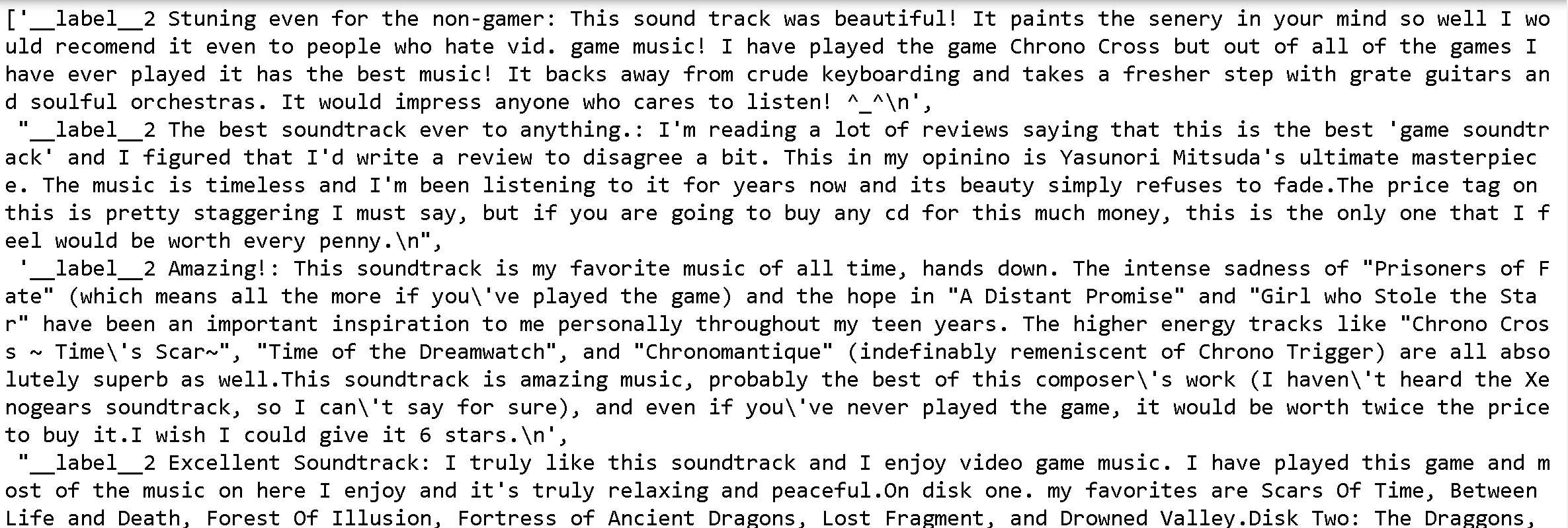
The two datasets were taken from the nltk library with each tweet having either a positive or negative sentiment. The positive tweet dataset and the negative tweet dataset needs to be combined and labeled before using a model to train it

Restaurant reviews-



This dataset looks at restaurant reviews by the customer with their appropriate sentiment labeled. This dataset was taken from Kaggle

Amazon reviews-



This dataset looks at amazon product reviews by the customer with their appropriate sentiment labeled. This dataset was taken from Kaggle

**Literature Review-**

The work with sentiment analysis [1] looks at comparing two models and the tradeoff between them in terms of computation and accuracy of the explainability. It didn’t dive deep into the datasets or the interpretability surrounding the dataset. [2] Looks into the need for explainable systems, compares different explainable systems, and brings up the need for a better explainable system for sentiment analysis in the field of AI. This paper also doesn’t deal with any datasets.

**Methodology-**

After pre-processing each of the datasets, they will be trained using a simple logistic regression classifier and a deep neural network for sentiment analysis. Afterward, their predictions will be analyzed using lime and captum. At first, looking at how each of the models works for each dataset and then comparing the models. The end goal is to create a pipeline that will output the main texts that go behind the sentiment and highlight the ones creating the inaccuracies with hopes of using the information to help in building better models, remove the ‘black-box’ notion of models for others, and make the dataset more understandable.

**References-**

<https://dacemirror.sci-hub.tw/proceedings-article/8d0bc00f1e268edfb14413481a2598bc/zucco2018.pdf?rand=5f67d6f42ab08?download=true>

[1] Zucco, C., Liang, H., Fatta, G. D., & Cannataro, M. (2018). *Explainable Sentiment Analysis with Applications in Medicine. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).* doi:10.1109/bibm.2018.8621359

<https://arxiv.org/pdf/1602.04938.pdf>

[2] Bodria, F., Panisson, A., Perotti, A., & Piaggesi, S. Explainability Methods for Natural Language Processing: Applications to Sentiment Analysis (Discussion Paper).

<https://www.aclweb.org/anthology/P19-1560.pdf>

[3] Liu, H., Yin, Q., & Wang, W. Y. (2018). Towards explainable NLP: A generative explanation framework for text classification. *arXiv preprint arXiv:1811.00196*.

<http://ceur-ws.org/Vol-2646/18-paper.pdf>

[4] Bodria, F., Panisson, A., Perotti, A., & Piaggesi, S. Explainability Methods for Natural Language Processing: Applications to Sentiment Analysis (Discussion Paper).

<https://towardsdatascience.com/rise-of-modern-nlp-and-the-need-of-interpretability-97dd4a655ac3>