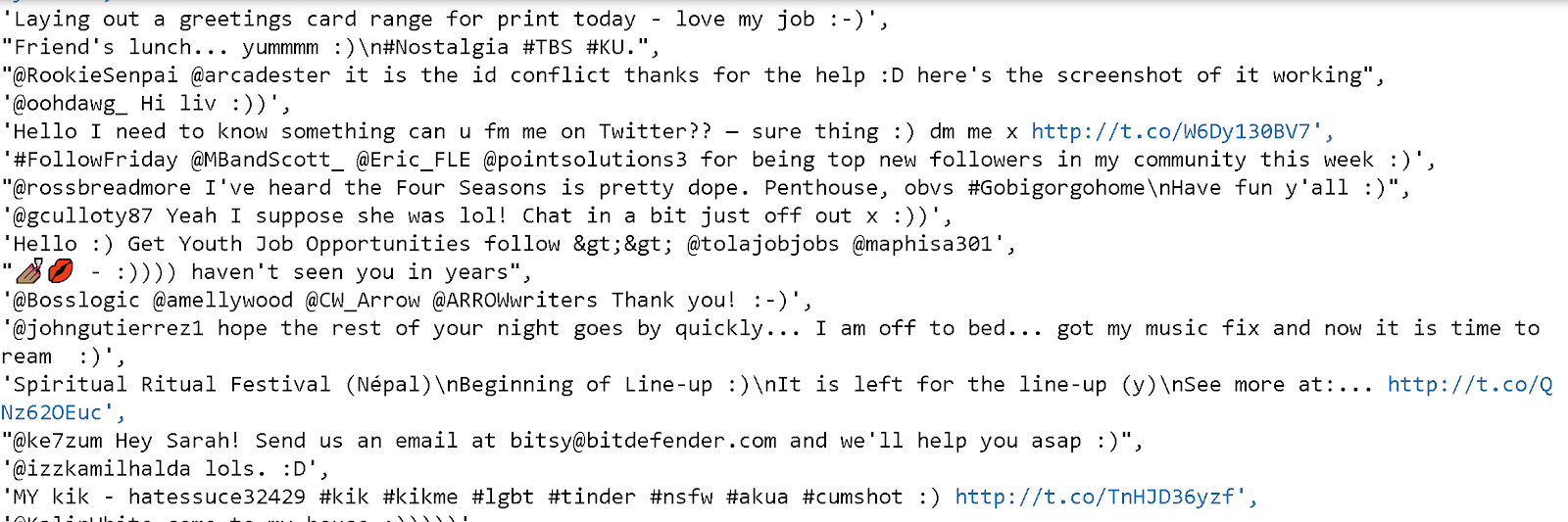
**Explainability of Transformer models in natural language processing for sentiment analysis**

**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 using transformer models to show what factors lead to the model’s prediction.

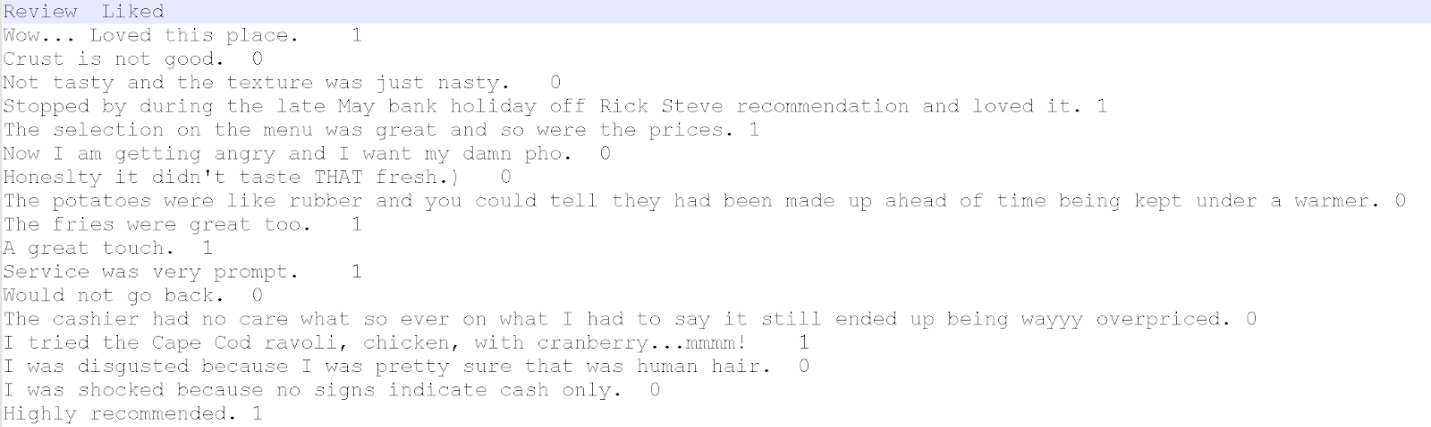
**Datasets-**

Positive and negative tweets:



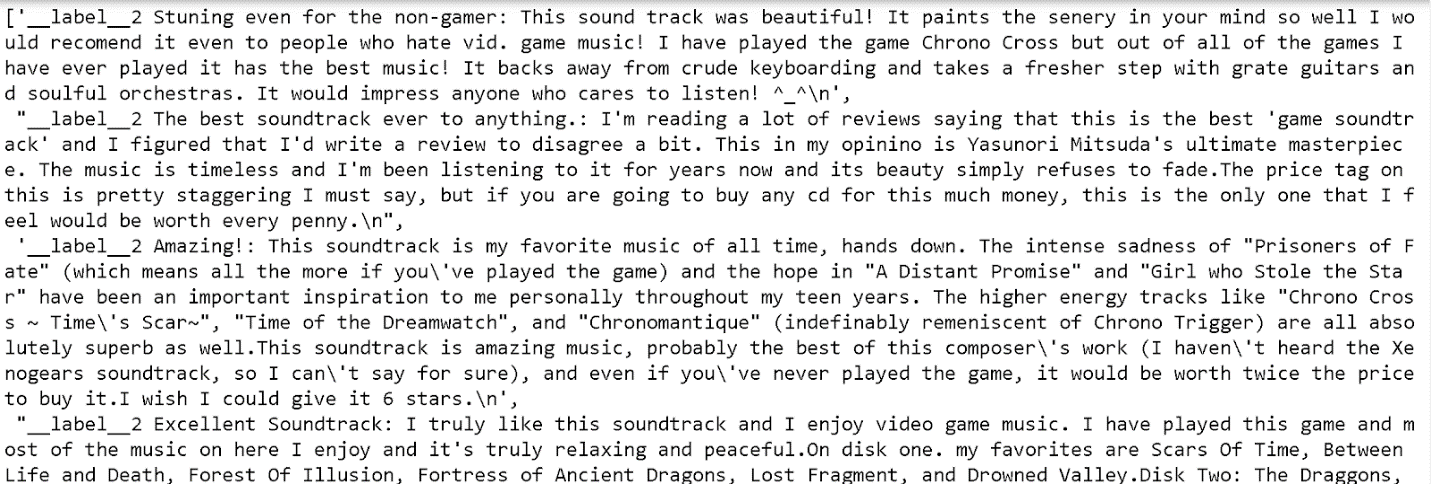
The two datasets were taken from the nltk library with each tweet having either a positive or negative sentiment. The positive tweet dataset and the negative tweet dataset needs to be combined and labeled before using a model to train it

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, a few transformer models are picked 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>

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

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

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

<http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

<https://arxiv.org/pdf/1810.04805.pdf>

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

<https://arxiv.org/pdf/1909.11942.pdf>

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

<https://arxiv.org/pdf/1907.11692.pdf>

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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

Abstract

Despite widespread adoption, machine learning models re-main mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one. In this work, we propose LIME, a novel explanation technique that explains the predictions of any classi\_er in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the exibility of these methods by explaining di\_erent models for text (e.g. random forests) and image classi\_cation (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classi\_er, and identifying why a classi\_er should not be trusted.

With so many popular Deep learning models being developed and solving complex issues, they still remain as black boxes. Many tools were created to work out the interpretability of such models in a few domains.

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 using transformer models to show what factors lead to the model’s prediction.

Introduction

Literature Review

Methodology

Results

Further Work

Conclusion