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:

```
'Laying out a greetings card range for print today - love my job :-)',
"Friend's lunch... yummmm :)\n#Nostalgia #TBS #KU.",
"@RookieSenpai @arcadester it is the id conflict thanks for the help :D here's the screenshot of it working",
'@oohdawg_ Hi liv :))',
'Hello I need to know something can u fm me on Twitter?? — sure thing :) dm me x http://t.co/W6Dy130BV7',
'#FollowFriday @MBandScott_ @Eric_FLE @pointsolutions3 for being top new followers in my community this week :)',
"@rossbreadmore I've heard the Four Seasons is pretty dope. Penthouse, obvs #Gobigorgohome\nHave fun y'all :)",
'@gculloty87 Yeah I suppose she was lol! Chat in a bit just off out x :))',
'Hello :) Get Youth Job Opportunities follow &gt;&gt; @tolajobjobs @maphisa301',
"💉💊 - :)))) haven't seen you in years",
'@Bosslogic @amellywood @CW_Arrow @ARROWwriters Thank you! :-)',
'@johngutierrez1 hope the rest of your night goes by quickly... I am off to bed... got my music fix and now it is time to
ream  :)',
'Spiritual Ritual Festival (Népal)\nBeginning of Line-up :)\nIt is left for the line-up (y)\nSee more at:... http://t.co/Q
Nz62OEuc',
"@ke7zum Hey Sarah! Send us an email at bitsy@bitdefender.com and we'll help you asap :)",
'@izzkamilhalda lols. :D',
'MY kik - hatessuce32429 #kik #kikme #lgbt #tinder #nsfw #akua #cumshot :) http://t.co/TnHJD36yzf',
```

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-

```
Review  Liked
Wow... Loved this place.    1
Crust is not good.  0
Not tasty and the texture was just nasty.   0
Stopped by during the late May bank holiday off Rick Steve recommendation and loved it. 1
The selection on the menu was great and so were the prices. 1
Now I am getting angry and I want my damn pho.  0
Honeslty it didn't taste THAT fresh.)   0
The potatoes were like rubber and you could tell they had been made up ahead of time being kept under a warmer. 0
The fries were great too.   1
A great touch.  1
Service was very prompt.    1
Would not go back.  0
The cashier had no care what so ever on what I had to say it still ended up being wayyy overpriced. 0
I tried the Cape Cod ravoli, chicken, with cranberry...mmmm!    1
I was disgusted because I was pretty sure that was human hair.  0
I was shocked because no signs indicate cash only.  0
Highly recommended. 1
```

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

Amazon reviews-

```
['__label__2 Stuning even for the non-gamer: This sound track was beautiful! It paints the senery in your mind so well I wo
uld recomend it even to people who hate vid. game music! I have played the game Chrono Cross but out of all of the games I
have ever played it has the best music! It backs away from crude keyboarding and takes a fresher step with grate guitars an
d soulful orchestras. It would impress anyone who cares to listen! ^_^\n',
 "__label__2 The best soundtrack ever to anything.: I'm reading a lot of reviews saying that this is the best 'game soundtr
ack' and I figured that I'd write a review to disagree a bit. This in my opinino is Yasunori Mitsuda's ultimate masterpiec
e. The music is timeless and I'm been listening to it for years now and its beauty simply refuses to fade.The price tag on
this is pretty staggering I must say, but if you are going to buy any cd for this much money, this is the only one that I f
eel would be worth every penny.\n",
 '__label__2 Amazing!: This soundtrack is my favorite music of all time, hands down. The intense sadness of "Prisoners of F
ate" (which means all the more if you\'ve played the game) and the hope in "A Distant Promise" and "Girl who Stole the Sta
r" have been an important inspiration to me personally throughout my teen years. The higher energy tracks like "Chrono Cros
s ~ Time\'s Scar~", "Time of the Dreamwatch", and "Chronomantique" (indefinably remeniscent of Chrono Trigger) are all abso
lutely superb as well.This soundtrack is amazing music, probably the best of this composer\'s work (I haven\'t heard the Xe
nogears soundtrack, so I can\'t say for sure), and even if you\'ve never played the game, it would be worth twice the price
to buy it.I wish I could give it 6 stars.\n',
 "__label__2 Excellent Soundtrack: I truly like this soundtrack and I enjoy video game music. I have played this game and m
ost of the music on here I enjoy and it's truly relaxing and peaceful.On disk one. my favorites are Scars Of Time, Between
Life and Death, Forest Of Illusion, Fortress of Ancient Dragons, Lost Fragment, and Drowned Valley.Disk Two: The Draggons,
```

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