# Final Project

## Minjie Yang

## 1. Introduction

Large language models (LLMs) have achieved significant success in tasks like financial sentiment analysis, where accurate predictions can directly influence decision-making processes. However, their complex architectures often make them challenging to interpret, raising concerns about their reliability and transparency in high-stakes domains such as finance. This report uses a BERT sentiment analysis model fine-tuned on financial news and explores its interpretability techniques, aiming to uncover its internal workings and enhance trust in its predictions.

The report is structured as follows:

- **Section 2** focuses on attention analysis, visualizing attention matrices to understand the model's focus on input features.
- **Section 3** investigates the impact of input perturbations on sentiment scores, analyzing the model's robustness.
- **Section 4** introduces loss sensitivity analysis, a novel approach to study how input changes affect the model’s loss function.

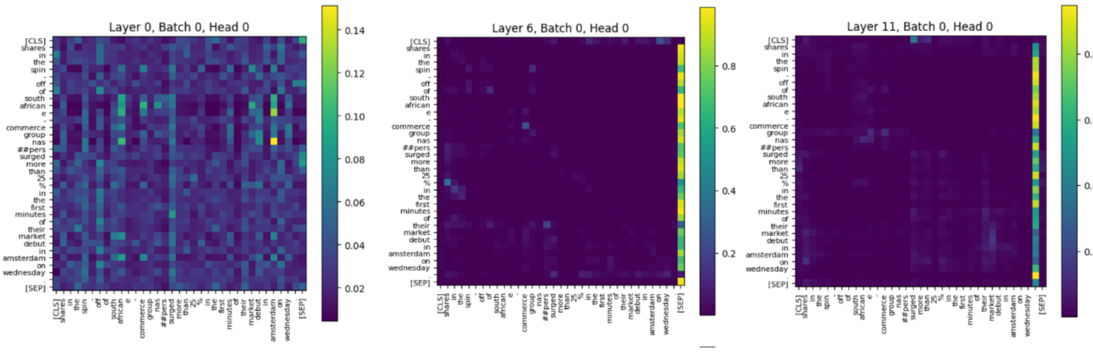
In each section, we analyze twenty positive and negative sentences respectively and identify the tokens that have the most significant impact on predictions based on different methods used in each section.

## 2. Attention Analysis

In this section, we obtained all attention matrices for each sample across different layers and heads. For both the positive and negative parts, we selected three specific layers for one sample from each to visualize. Subsequently, we summed the attention scores corresponding to each token, identifying the tokens with the highest attention scores as those most influential to the prediction (excluding irrelevant tokens such as CLS, SEP, and punctuation marks).

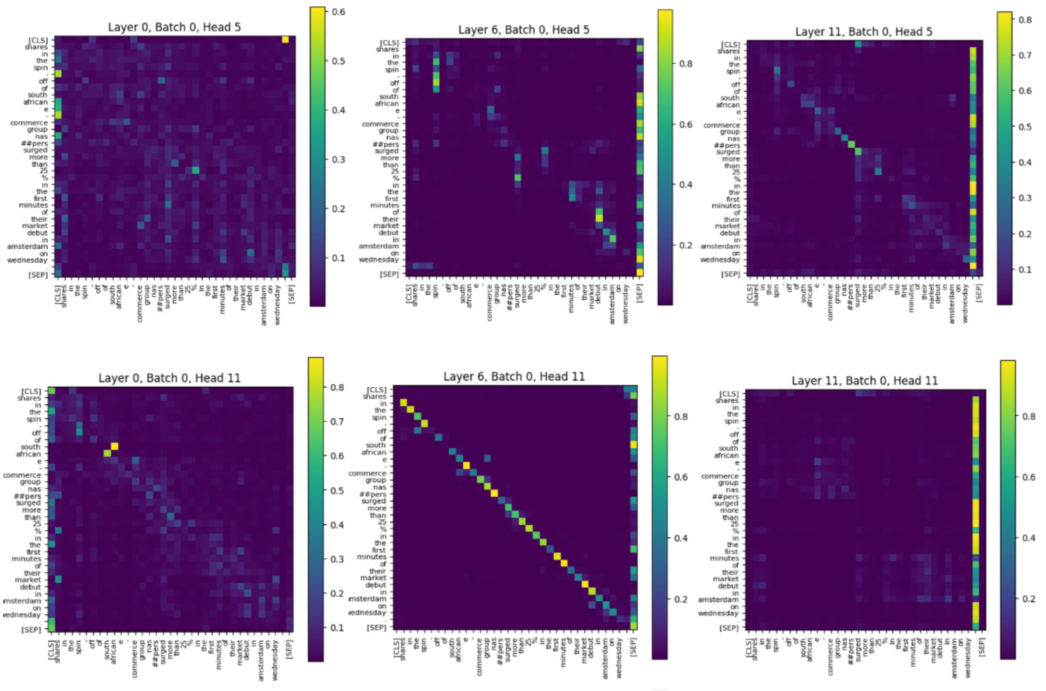
### 2.1 Positive Part

We randomly selected a sentence predicted as positive ("Shares in the spin-off of South African e-commerce group Naspers surged more than 25% in the first minutes of their market debut in Amsterdam on Wednesday.") to visualize the attention of all heads in its 1st, 7th, and 12th layers. In the report, we showcased a few representative heads.


Layer 0, Batch 0, Head 0

Layer 6, Batch 0, Head 0

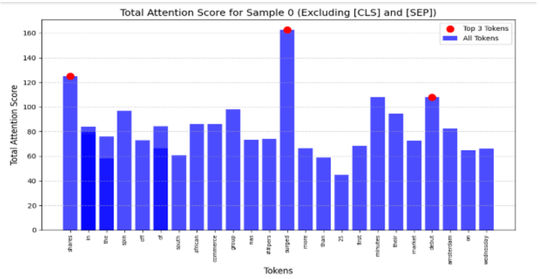
Layer 11, Batch 0, Head 0

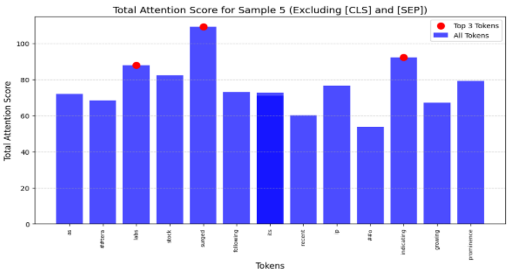


From the figures, we can observe clear differences in the attention patterns across layers and heads. In the initial layers, the attention matrices are relatively dispersed, with no specific tokens receiving noticeably higher attention weights. This suggests that the model is in the early stages of feature extraction, where attention is distributed across all tokens to capture broad contextual information. By the 7th layer, we see a shift in focus, with the [SEP] token attracting significantly higher attention. This indicates that the model begins using [SEP] as a boundary or aggregation token, potentially as a marker for summarizing the semantic content of the sentence. In the final layer, the attention becomes highly concentrated, with the [.] token (punctuation) receiving the majority of the attention. This behavior suggests that the model might be leveraging punctuation as a key delimiter for finalizing sentence-level representations or as a structural cue in its prediction process.

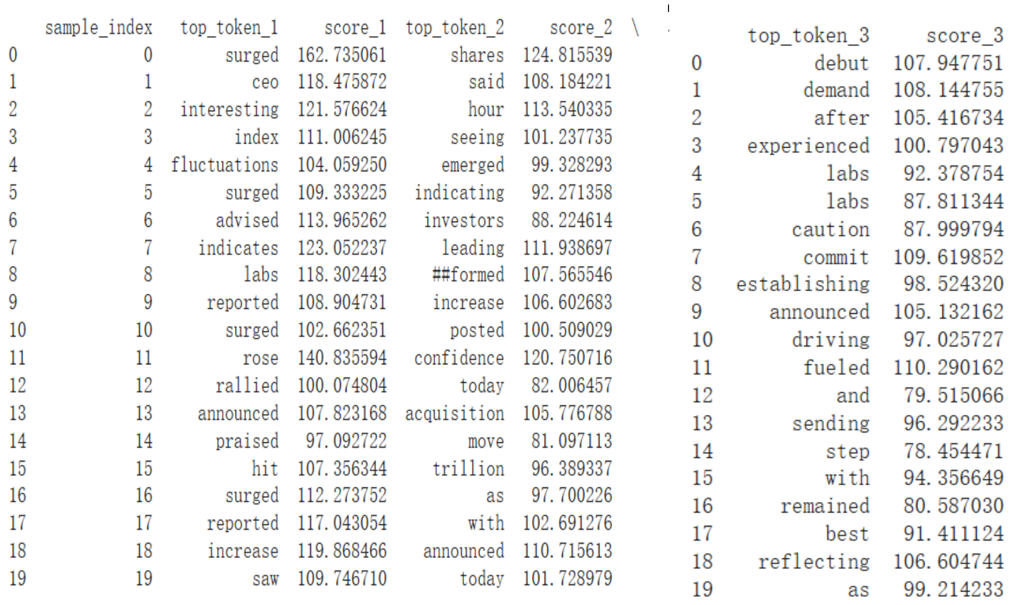
When analyzing the behavior across different heads, the first and last layers of each head appear relatively similar, likely reflecting consistent initial and final processing patterns across the model. However, in the middle layers, significant differences emerge between heads. For instance, in the 7th layer, the first head and last head allocate almost all their attention to the [SEP] token, while other tokens receive minimal attention. This indicates a specialized role for these heads, possibly focusing on sentence segmentation or aggregation tasks. In contrast, the 6th head in the same layer not only assigns high attention to the [SEP] token but also shows elevated attention for each token to itself. This self-attention behavior suggests that this head might be refining individual token representations by reinforcing their unique contributions to the overall sentence structure.

Next, I summed the attention scores of different tokens across different heads and layers, and then calculated the top 3 tokens for each sample based on their scores.





Hint: In this analysis, we excluded punctuation tokens as well as the special tokens [CLS] and [SEP], focusing only on the attention scores assigned to the meaningful tokens within the sentence. The figures display the aggregated attention scores for different tokens across two samples, showcasing which words the model deemed most important. Notably, tokens such as "surged," "debut," and "indicating" stood out with the highest total attention scores. This suggests that these words play a critical role in how the model interprets and understands the sentence.

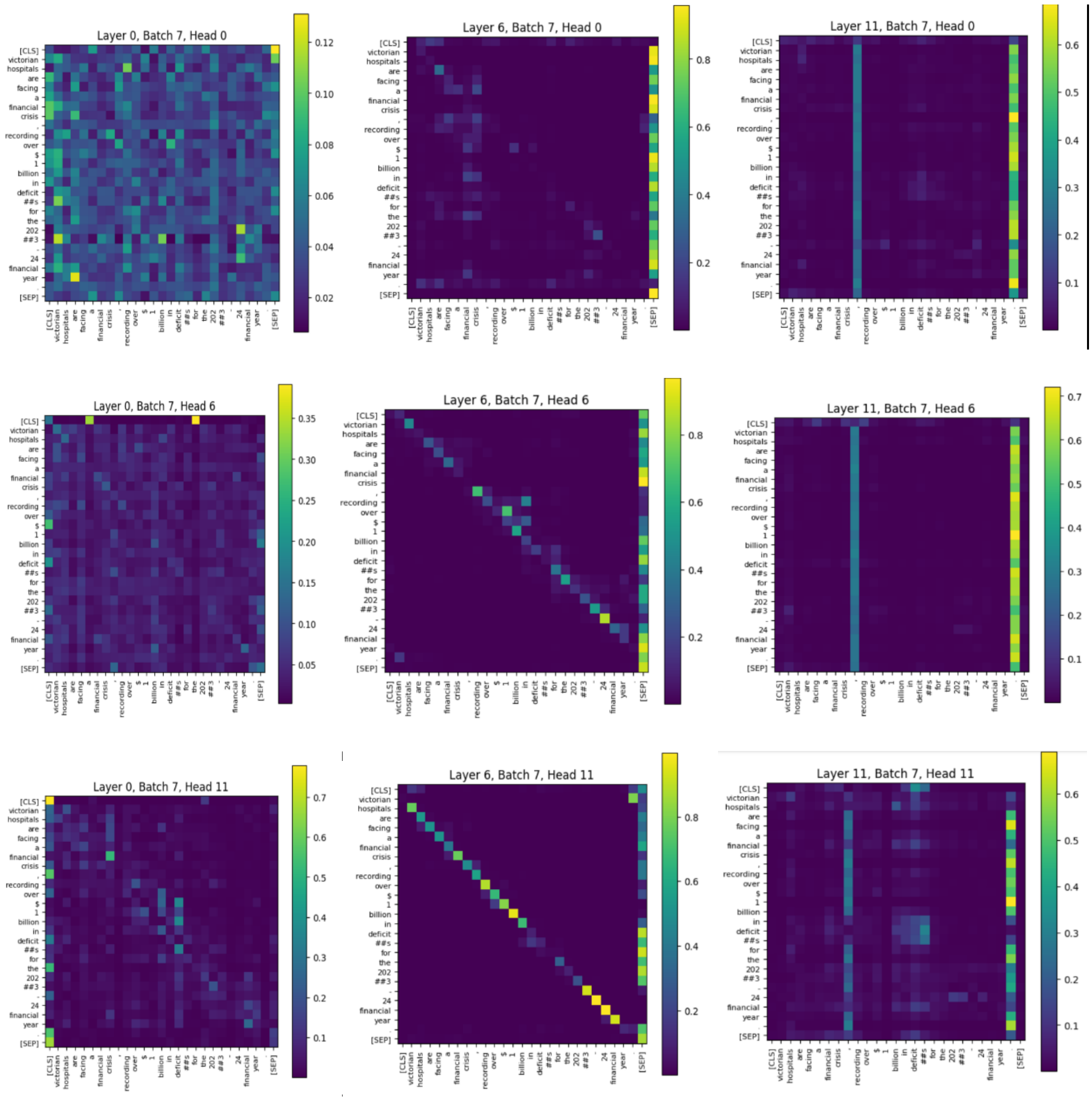

```
    sample_index  top_token_1     score_1  top_token_2     score_2  \
0              0       surged  162.735061       shares  124.815539
1              1          ceo  118.475872         said  108.184221
2              2  interesting  121.576624         hour  113.540335
3              3        index  111.006245       seeing  101.237735
4              4  fluctuations  104.059250      emerged   99.328293
5              5       surged  109.333225   indicating   92.271358
6              6      advised  113.965262    investors   88.224614
7              7    indicates  123.052237      leading  111.938697
8              8         labs  118.302443    ##formed  107.565546
9              9     reported  108.904731     increase  106.602683
10            10       surged  102.662351       posted  100.509029
11            11         rose  140.835594   confidence  120.750716
12            12      rallied  100.074804        today   82.006457
13            13    announced  107.823168  acquisition  105.776788
14            14      praised   97.092722         move   81.097113
15            15          hit  107.356344     trillion   96.389337
16            16       surged  112.273752           as   97.700226
17            17     reported  117.043054         with  102.691276
18            18     increase  119.868466    announced  110.715613
19            19          saw  109.746710        today  101.728979

     top_token_3     score_3
0          debut  107.947751
1         demand  108.144755
2          after  105.416734
3    experienced  100.797043
4           labs   92.378754
5           labs   87.811344
6        caution   87.999794
7         commit  109.619852
8   establishing   98.524320
9      announced  105.132162
10       driving   97.025727
11        fueled  110.290162
12           and   79.515066
13       sending   96.292233
14          step   78.454471
15          with   94.356649
16      remained   80.587030
17          best   91.411124
18    reflecting  106.604744
19            as   99.214233
```

The above figures show the top 3 tokens with the highest attention scores across all test samples. We can observe that tokens such as "surged," "interesting," "rose," "confidence," and "shares" all have attention scores exceeding 120.

This indicates that these tokens play a significant role in the model's interpretation and decision-making process. Words like "surged" and "rose" are action-oriented, often associated with positive sentiment, especially in financial contexts, which aligns with the FinBERT model's focus on financial text analysis. Similarly, "confidence" and "shares" are key domain-specific terms that frequently appear in financial discourse, reflecting the importance of understanding market sentiment and activities. The inclusion of "interesting" as a highly attended token suggests that the model may also give weight to words that provide subjective or evaluative context, further influencing the sentiment classification.

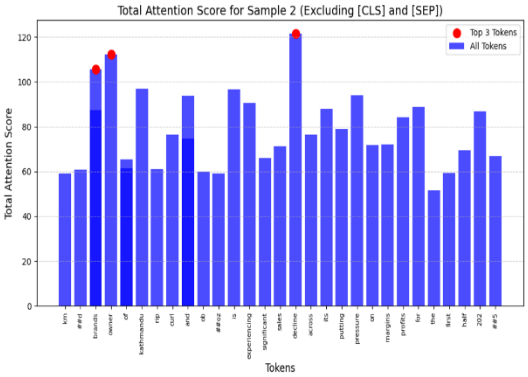
## 2.2 Negative Part

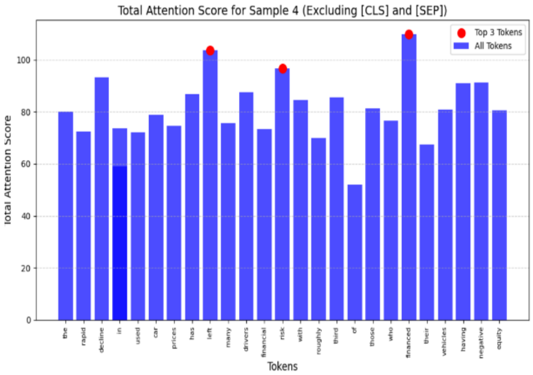
Here, we selected a sample predicted as negative: "Victorian hospitals are facing a financial crisis, recording over $1 billion in deficits for the 2023-24 financial year."



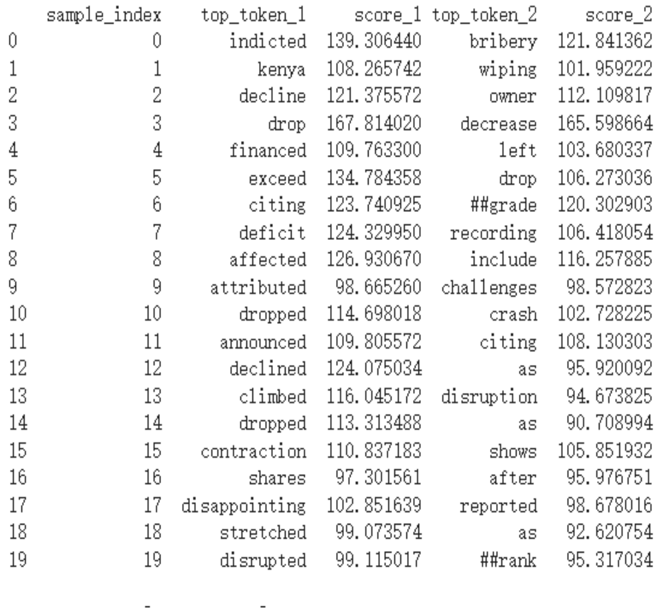
The attention pattern here is similar to that observed in the positive sample. In Layer 1, the attention matrix is relatively dispersed, with most tokens receiving low attention scores. However, as the layers progress, the majority of the attention becomes concentrated on punctuation and the [SEP] token. In the middle layers, attention matrices from the later heads show a notable pattern where different tokens assign high attention to themselves.

This progression reflects the model's typical behavior during hierarchical representation learning. In the initial layers, attention is distributed broadly, allowing the model to capture general contextual relationships across all tokens. This stage serves as a foundation for deeper semantic extraction. In the middle layers, the model begins refining these representations, as seen in the heads where tokens focus more on themselves. This self-attention mechanism likely helps the model enhance individual token embeddings, emphasizing their specific contributions within the sentence context. Finally, in the later layers, the attention converges on punctuation and the [SEP] token. This behavior suggests that these tokens act as aggregation points, consolidating the information necessary for the model to make predictions.

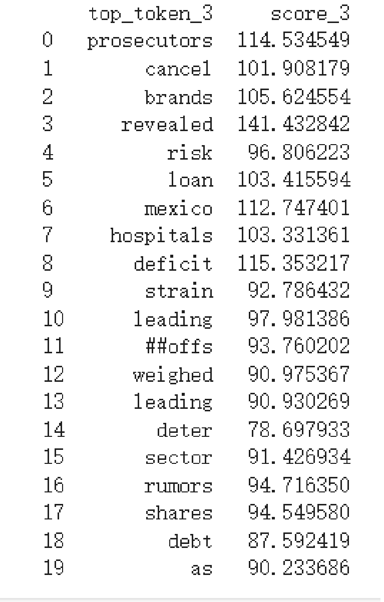




In the two samples shown in the figure, we observe that tokens such as "financed," "left," and "declined" have notably high attention scores. These tokens likely play a critical role in how the model interprets and classifies the sentences. For instance, "declined" conveys a clear negative sentiment, which is essential for identifying the overall tone of the sentence, especially in financial or evaluative contexts. Similarly, "financed" and "left" might act as key indicators of contextual shifts or significant events, providing the model with important clues about the narrative's direction.



```
    sample_index   top_token_1     score_1 top_token_2     score_2
0              0      indicted  139.306440     bribery  121.841362
1              1         kenya  108.265742      wiping  101.959222
2              2       decline  121.375572       owner  112.109817
3              3          drop  167.814020    decrease  165.598664
4              4      financed  109.763300        left  103.680337
5              5        exceed  134.784358        drop  106.273036
6              6        citing  123.740925     ##grade  120.302903
7              7       deficit  124.329950   recording  106.418054
8              8      affected  126.930670     include  116.257885
9              9    attributed   98.665260  challenges   98.572823
10            10       dropped  114.698018       crash  102.728225
11            11     announced  109.805572      citing  108.130303
12            12      declined  124.075034          as   95.920092
13            13       climbed  116.045172  disruption   94.673825
14            14       dropped  113.313488          as   90.708994
15            15   contraction  110.837183       shows  105.851932
16            16        shares   97.301561       after   95.976751
17            17 disappointing  102.851639    reported   98.678016
18            18     stretched   99.073574          as   92.620754
19            19     disrupted   99.115017      ##rank   95.317034
```



```
     top_token_3     score_3
0    prosecutors  114.534549
1         cancel  101.908179
2         brands  105.624554
3       revealed  141.432842
4           risk   96.806223
5           loan  103.415594
6         mexico  112.747401
7      hospitals  103.331361
8        deficit  115.353217
9         strain   92.786432
10       leading   97.981386
11        ##offs   93.760202
12       weighed   90.975367
13       leading   90.930269
14         deter   78.697933
15        sector   91.426934
16        rumors   94.716350
17        shares   94.549580
18          debt   87.592419
19            as   90.233686
```

Among the top 3 tokens, we observe that "indicated," "decline," "drop," "exceed," "citing," "deficit," "affected," "bribery," "decrease," and "revealed" all have total attention scores exceeding 120. This suggests that these words play a significant role in shaping the model's negative predictions. Some of these tokens, such as "decline," "drop," "deficit," and "decrease," are directly associated with negative sentiment and align well with the expected behavior of the model. These words often appear in contexts describing loss, reduction, or unfavorable outcomes, making their high attention scores logical and aligned with the overall negative prediction.

However, the prominence of certain other tokens, like "indicated" and "exceed," may seem less intuitive. While they can be associated with neutral or even positive contexts depending on usage, their high attention scores here might indicate biases in the training data or the model's tendency to overemphasize specific words without fully contextualizing their meanings.

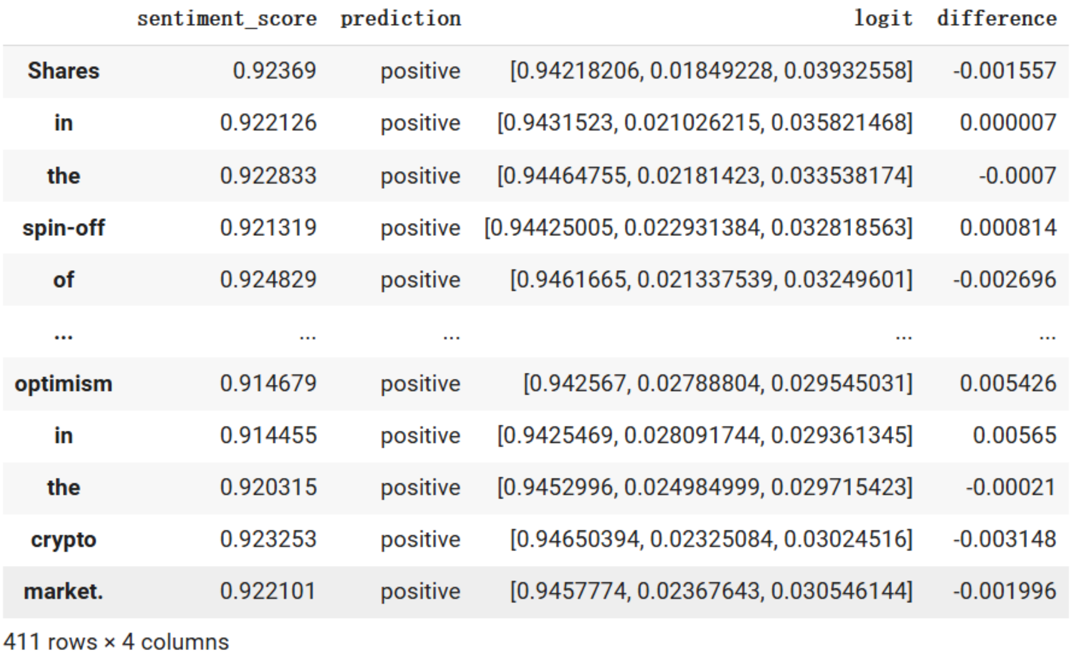
# 3. Input Perturbation based on sentiment score

In this section, we evaluate the importance of each word in a sentence by individually removing each token and observing its impact on the sentiment score. By comparing the sentiment scores before and after the removal of a word, we are able to measure the contribution of that specific token to the overall prediction. This approach helps us identify the tokens that are most influential in driving the model's decision-making process.

The sentiment score of this model is calculated by subtracting the logit of the negative prediction from the logit of the positive prediction. Therefore, a larger positive sentiment score indicates a higher confidence in the model's positive prediction, while a smaller negative sentiment score indicates a higher confidence in its negative prediction.

## 3.1 Positive Part

We deleted each word from 20 sentences predicted as positive and re-evaluated the predictions, resulting in the figure shown below.

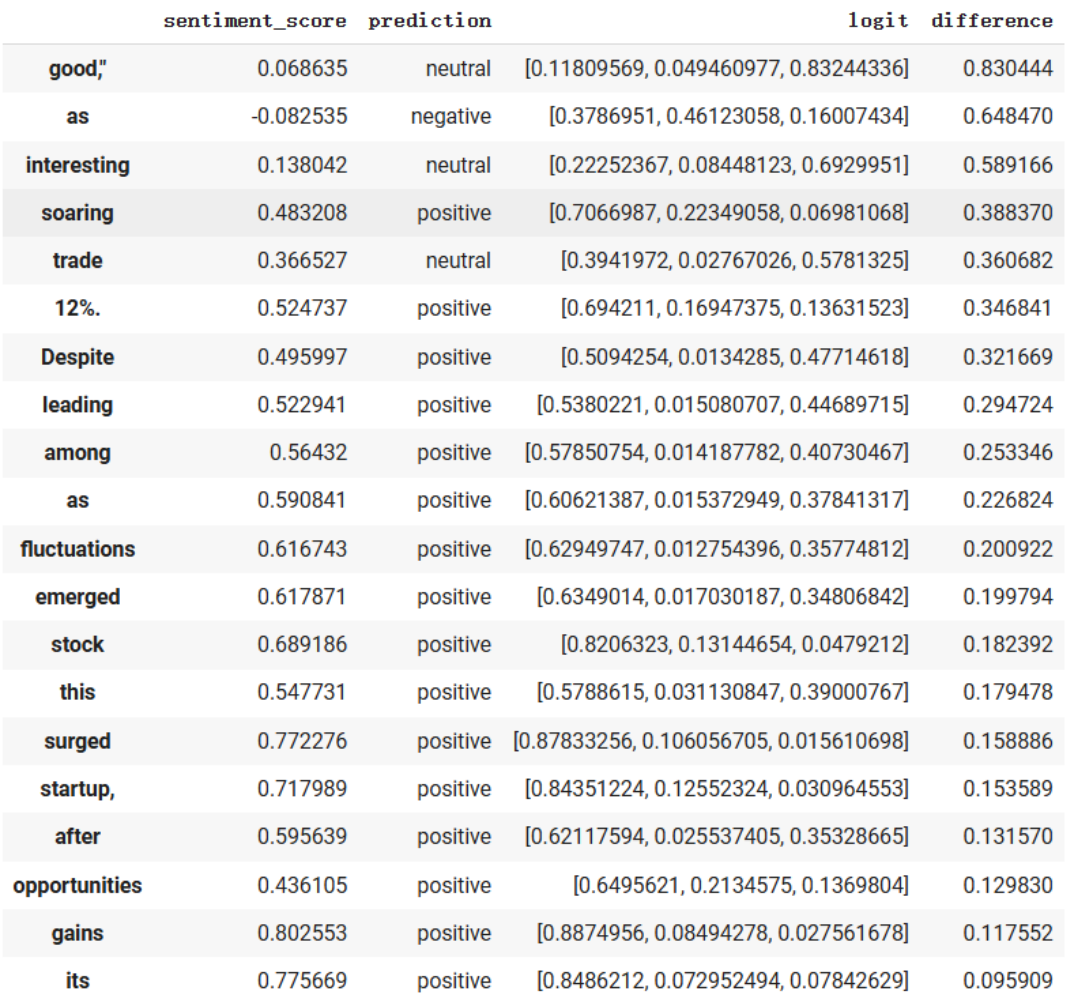


| | sentiment_score | prediction | logit | difference |
|---|---|---|---|---|
| **Shares** | 0.92369 | positive | [0.94218206, 0.01849228, 0.03932558] | -0.001557 |
| **in** | 0.922126 | positive | [0.9431523, 0.021026215, 0.035821468] | 0.000007 |
| **the** | 0.922833 | positive | [0.94464755, 0.02181423, 0.033538174] | -0.0007 |
| **spin-off** | 0.921319 | positive | [0.94425005, 0.022931384, 0.032818563] | 0.000814 |
| **of** | 0.924829 | positive | [0.9461665, 0.021337539, 0.03249601] | -0.002696 |
| **...** | ... | ... | ... | ... |
| **optimism** | 0.914679 | positive | [0.942567, 0.02788804, 0.029545031] | 0.005426 |
| **in** | 0.914455 | positive | [0.9425469, 0.028091744, 0.029361345] | 0.00565 |
| **the** | 0.920315 | positive | [0.9452996, 0.024984999, 0.029715423] | -0.00021 |
| **crypto** | 0.923253 | positive | [0.94650394, 0.02325084, 0.03024516] | -0.003148 |
| **market.** | 0.922101 | positive | [0.9457774, 0.02367643, 0.030546144] | -0.001996 |

411 rows × 4 columns

The difference column is calculated by subtracting the sentiment score of the new sample (after deleting a word) from the sentiment score of the original sample. From the figure, we can see that the deletion of most words has little impact on the model's predictions, indicating that these words are not critical to the model's sentiment classification

Next, we identified and analyzed the 20 words with the largest differences, as shown in the figure below. These words had the greatest impact on the sentiment score when deleted, indicating their critical importance in the model's sentiment predictions.



| | sentiment_score | prediction | logit | difference |
|---|---|---|---|---|
| **good,"** | 0.068635 | neutral | [0.11809569, 0.049460977, 0.83244336] | 0.830444 |
| **as** | -0.082535 | negative | [0.3786951, 0.46123058, 0.16007434] | 0.648470 |
| **interesting** | 0.138042 | neutral | [0.22252367, 0.08448123, 0.6929951] | 0.589166 |
| **soaring** | 0.483208 | positive | [0.7066987, 0.22349058, 0.06981068] | 0.388370 |
| **trade** | 0.366527 | neutral | [0.3941972, 0.02767026, 0.5781325] | 0.360682 |
| **12%.** | 0.524737 | positive | [0.694211, 0.16947375, 0.13631523] | 0.346841 |
| **Despite** | 0.495997 | positive | [0.5094254, 0.0134285, 0.47714618] | 0.321669 |
| **leading** | 0.522941 | positive | [0.5380221, 0.015080707, 0.44689715] | 0.294724 |
| **among** | 0.56432 | positive | [0.57850754, 0.014187782, 0.40730467] | 0.253346 |
| **as** | 0.590841 | positive | [0.60621387, 0.015372949, 0.37841317] | 0.226824 |
| **fluctuations** | 0.616743 | positive | [0.62949747, 0.012754396, 0.35774812] | 0.200922 |
| **emerged** | 0.617871 | positive | [0.6349014, 0.017030187, 0.34806842] | 0.199794 |
| **stock** | 0.689186 | positive | [0.8206323, 0.13144654, 0.0479212] | 0.182392 |
| **this** | 0.547731 | positive | [0.5788615, 0.031130847, 0.39000767] | 0.179478 |
| **surged** | 0.772276 | positive | [0.87833256, 0.106056705, 0.015610698] | 0.158886 |
| **startup,** | 0.717989 | positive | [0.84351224, 0.12552324, 0.030964553] | 0.153589 |
| **after** | 0.595639 | positive | [0.62117594, 0.025537405, 0.35328665] | 0.131570 |
| **opportunities** | 0.436105 | positive | [0.6495621, 0.2134575, 0.1369804] | 0.129830 |
| **gains** | 0.802553 | positive | [0.8874956, 0.08494278, 0.027561678] | 0.117552 |
| **its** | 0.775669 | positive | [0.8486212, 0.072952494, 0.07842629] | 0.095909 |

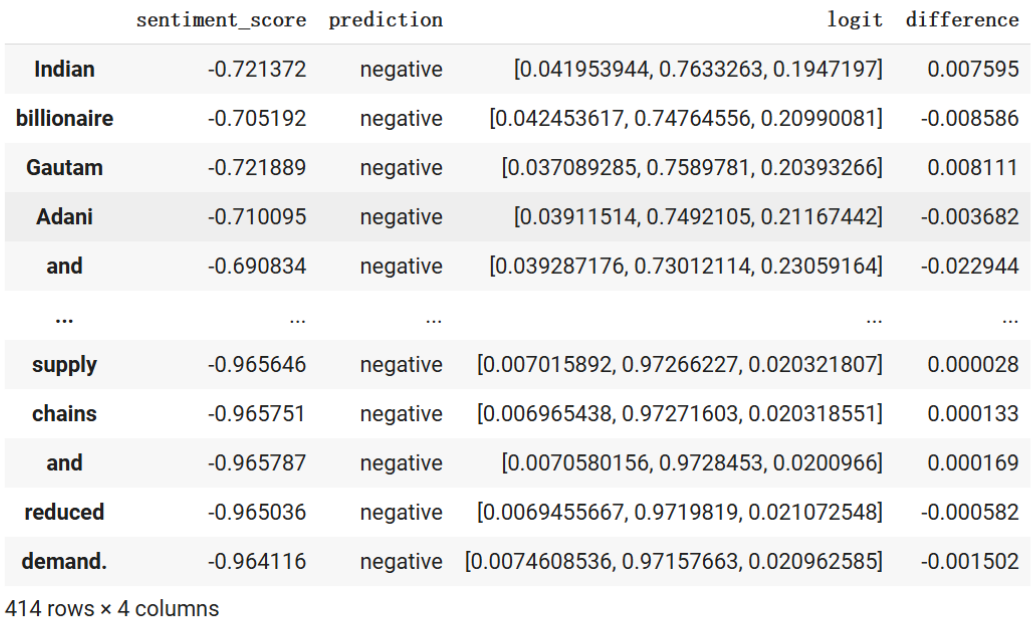
The table highlights the 20 most influential tokens in terms of sentiment score differences when removed from their respective sentences. Tokens such as *"good,"*, *"soaring"*, and *"gains"* demonstrate significant importance in the model's sentiment predictions, with *"good,"* exhibiting the largest logit difference (0.830444). These words are highly domain-relevant, particularly in the financial context, where terms like *"soaring"* and *"gains"* strongly align with positive sentiment predictions.

Interestingly, some tokens such as *"as"* and *"Despite"* show substantial influence despite their less direct connection to sentiment, suggesting the model may rely on contextual or structural roles these tokens play in shaping the overall sentence meaning.

Moreover, words like *"good,"* and *"interesting"* originally contributed to positive predictions, but their deletion caused the model to reclassify the samples as neutral. This demonstrates that these words are not only highly influential but also critical for the model to maintain its original sentiment classification. Such shifts suggest that the model heavily depends on specific tokens to anchor its predictions, which can lead to vulnerabilities in cases where these tokens are absent or misinterpreted.

## 3.2 Negative Part

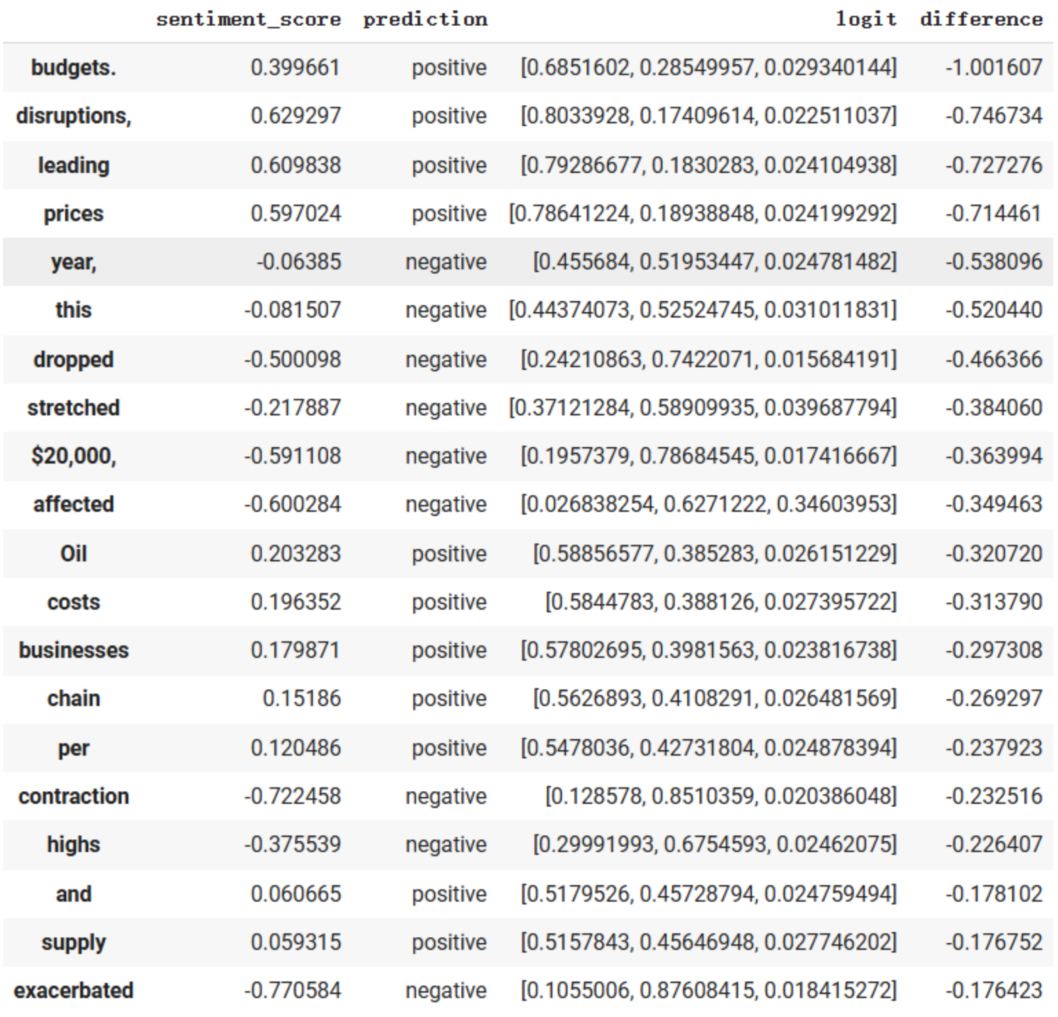
We deleted each word from 20 sentences predicted as negative and re-evaluated the predictions, resulting in the figure shown below.



| | sentiment_score | prediction | logit | difference |
|---|---|---|---|---|
| **Indian** | -0.721372 | negative | [0.041953944, 0.7633263, 0.1947197] | 0.007595 |
| **billionaire** | -0.705192 | negative | [0.042453617, 0.74764556, 0.20990081] | -0.008586 |
| **Gautam** | -0.721889 | negative | [0.037089285, 0.7589781, 0.20393266] | 0.008111 |
| **Adani** | -0.710095 | negative | [0.03911514, 0.7492105, 0.21167442] | -0.003682 |
| **and** | -0.690834 | negative | [0.039287176, 0.73012114, 0.23059164] | -0.022944 |
| **...** | ... | ... | ... | ... |
| **supply** | -0.965646 | negative | [0.007015892, 0.97266227, 0.020321807] | 0.000028 |
| **chains** | -0.965751 | negative | [0.006965438, 0.97271603, 0.020318551] | 0.000133 |
| **and** | -0.965787 | negative | [0.0070580156, 0.9728453, 0.0200966] | 0.000169 |
| **reduced** | -0.965036 | negative | [0.0069455667, 0.9719819, 0.021072548] | -0.000582 |
| **demand.** | -0.964116 | negative | [0.0074608536, 0.97157663, 0.020962585] | -0.001502 |

414 rows × 4 columns

This figure demonstrates a pattern similar to the previous one.



| | sentiment_score | prediction | logit | difference |
|---|---|---|---|---|
| **budgets.** | 0.399661 | positive | [0.6851602, 0.28549957, 0.029340144] | -1.001607 |
| **disruptions,** | 0.629297 | positive | [0.8033928, 0.17409614, 0.022511037] | -0.746734 |
| **leading** | 0.609838 | positive | [0.79286677, 0.1830283, 0.024104938] | -0.727276 |
| **prices** | 0.597024 | positive | [0.78641224, 0.18938848, 0.024199292] | -0.714461 |
| **year,** | -0.06385 | negative | [0.455684, 0.51953447, 0.024781482] | -0.538096 |
| **this** | -0.081507 | negative | [0.44374073, 0.52524745, 0.031011831] | -0.520440 |
| **dropped** | -0.500098 | negative | [0.24210863, 0.7422071, 0.015684191] | -0.466366 |
| **stretched** | -0.217887 | negative | [0.37121284, 0.58909935, 0.039687794] | -0.384060 |
| **$20,000,** | -0.591108 | negative | [0.1957379, 0.78684545, 0.017416667] | -0.363994 |
| **affected** | -0.600284 | negative | [0.026838254, 0.6271222, 0.34603953] | -0.349463 |
| **Oil** | 0.203283 | positive | [0.58856577, 0.385283, 0.026151229] | -0.320720 |
| **costs** | 0.196352 | positive | [0.5844783, 0.388126, 0.027395722] | -0.313790 |
| **businesses** | 0.179871 | positive | [0.57802695, 0.3981563, 0.023816738] | -0.297308 |
| **chain** | 0.15186 | positive | [0.5626893, 0.4108291, 0.026481569] | -0.269297 |
| **per** | 0.120486 | positive | [0.5478036, 0.42731804, 0.024878394] | -0.237923 |
| **contraction** | -0.722458 | negative | [0.128578, 0.8510359, 0.020386048] | -0.232516 |
| **highs** | -0.375539 | negative | [0.29991993, 0.6754593, 0.02462075] | -0.226407 |
| **and** | 0.060665 | positive | [0.5179526, 0.45728794, 0.024759494] | -0.178102 |
| **supply** | 0.059315 | positive | [0.5157843, 0.45646948, 0.027746202] | -0.176752 |
| **exacerbated** | -0.770584 | negative | [0.1055006, 0.87608415, 0.018415272] | -0.176423 |

This table analyzes top20 tokens influencing prediction. Tokens such as *"budgets,"*, *"disruptions,"*, and *"leading"* show significant logit differences, indicating their critical role in maintaining negative sentiment predictions. Domain-specific terms like *"prices,"*, *"$20,000,"*, and *"businesses"* heavily influence the model's understanding, reflecting their importance in financial contexts where they often convey evaluative meaning.

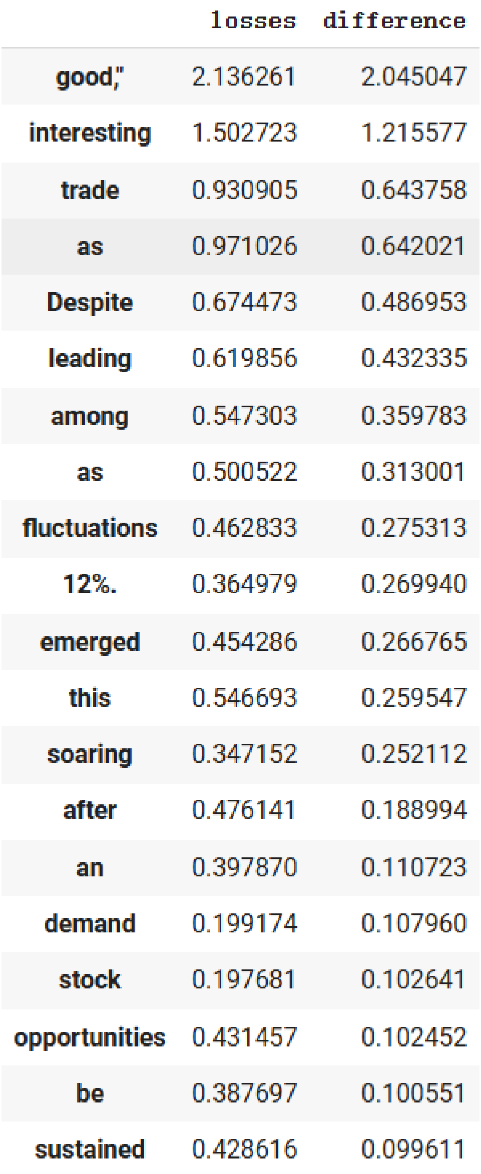
Interestingly, some less sentiment-laden tokens, such as *"and"* and *"per,"*, also show influence, likely due to their structural importance within sentences. Additionally, tokens strongly associated with negative sentiment, such as *"contraction,"*, *"exacerbated,"*, and *"dropped,"* consistently anchor the model's predictions, maintaining negative classifications even with minor adjustments. Overall, the analysis underscores the model's reliance on both sentiment-heavy and contextual tokens, while also revealing areas where certain structural words disproportionately affect predictions.

# 4. Input Perturbation based on loss

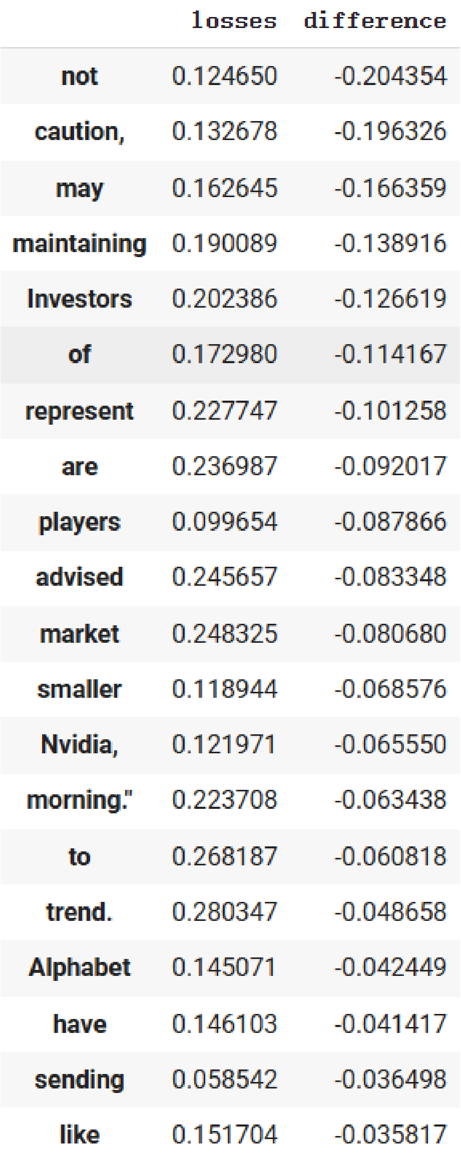
In this section, we evaluate the impact of each word in a sentence by removing tokens one at a time and examining the resulting change in the loss function. The difference in loss before and after the removal highlights the importance of the token in shaping the prediction.

The overall process is similar to the third part, except that the sentiment score is replaced with the loss. However, when analyzing the top 20 tokens here, we considered not only the top 20 highest values but also the top 20 lowest values.

## 4.1 Positive Part



| | losses | difference |
|---|---|---|
| good," | 2.136261 | 2.045047 |
| interesting | 1.502723 | 1.215577 |
| trade | 0.930905 | 0.643758 |
| as | 0.971026 | 0.642021 |
| Despite | 0.674473 | 0.486953 |
| leading | 0.619856 | 0.432335 |
| among | 0.547303 | 0.359783 |
| as | 0.500522 | 0.313001 |
| fluctuations | 0.462833 | 0.275313 |
| 12%. | 0.364979 | 0.269940 |
| emerged | 0.454286 | 0.266765 |
| this | 0.546693 | 0.259547 |
| soaring | 0.347152 | 0.252112 |
| after | 0.476141 | 0.188994 |
| an | 0.397870 | 0.110723 |
| demand | 0.199174 | 0.107960 |
| stock | 0.197681 | 0.102641 |
| opportunities | 0.431457 | 0.102452 |
| be | 0.387697 | 0.100551 |
| sustained | 0.428616 | 0.099611 |

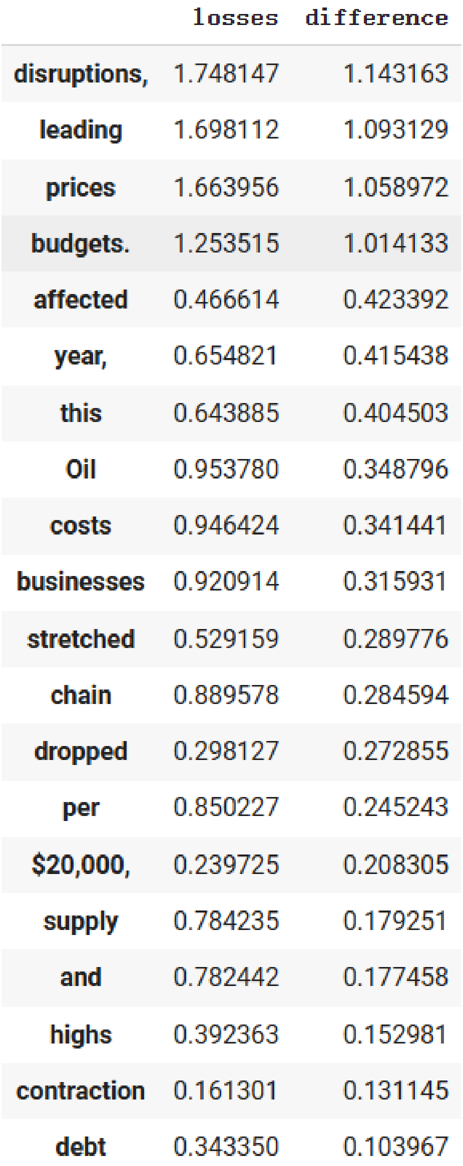


| | losses | difference |
|---|---|---|
| not | 0.124650 | -0.204354 |
| caution, | 0.132678 | -0.196326 |
| may | 0.162645 | -0.166359 |
| maintaining | 0.190089 | -0.138916 |
| Investors | 0.202386 | -0.126619 |
| of | 0.172980 | -0.114167 |
| represent | 0.227747 | -0.101258 |
| are | 0.236987 | -0.092017 |
| players | 0.099654 | -0.087866 |
| advised | 0.245657 | -0.083348 |
| market | 0.248325 | -0.080680 |
| smaller | 0.118944 | -0.068576 |
| Nvidia, | 0.121971 | -0.065550 |
| morning." | 0.223708 | -0.063438 |
| to | 0.268187 | -0.060818 |
| trend. | 0.280347 | -0.048658 |
| Alphabet | 0.145071 | -0.042449 |
| have | 0.146103 | -0.041417 |
| sending | 0.058542 | -0.036498 |
| like | 0.151704 | -0.035817 |

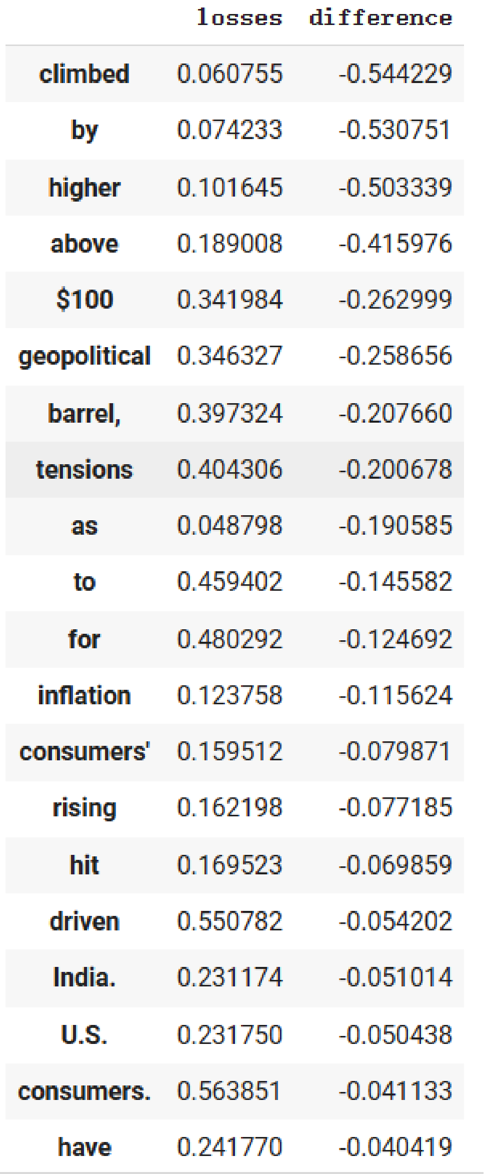
The left figure shows the top 20 highest values, while the right figure shows the top 20 lowest values. The left table highlights the top 20 tokens with the largest differences in loss values when removed, showcasing their significant impact on the model's performance. Tokens such as *"good,"*, *"interesting"*, and *"trade"* have the highest differences, indicating their critical role in shaping the model's predictions. Many of these tokens, like *"good,"*, *"soaring"*, and *"opportunities"*, are strongly associated with positive sentiment, suggesting the model heavily relies on these words to capture and reinforce positive outcomes. Additionally, the inclusion of contextually important tokens like *"Despite"* and *"fluctuations"* reflects the model's ability to consider nuanced sentence structures and contextual cues.

The right table showcases the 20 tokens with the smallest differences in loss values when removed, revealing that these tokens not only contribute minimally to the model's predictions but may also negatively impact its original predictions. Tokens such as *"not,"*, *"caution,"*, and *"may"* exhibit the largest negative differences, suggesting that their presence slightly increases the loss, indicating they could be introducing noise or ambiguity into the model's decision-making process.

## 4.2 Negative Part



| | losses | difference |
|---|---|---|
| disruptions, | 1.748147 | 1.143163 |
| leading | 1.698112 | 1.093129 |
| prices | 1.663956 | 1.058972 |
| budgets. | 1.253515 | 1.014133 |
| affected | 0.466614 | 0.423392 |
| year, | 0.654821 | 0.415438 |
| this | 0.643885 | 0.404503 |
| Oil | 0.953780 | 0.348796 |
| costs | 0.946424 | 0.341441 |
| businesses | 0.920914 | 0.315931 |
| stretched | 0.529159 | 0.289776 |
| chain | 0.889578 | 0.284594 |
| dropped | 0.298127 | 0.272855 |
| per | 0.850227 | 0.245243 |
| $20,000, | 0.239725 | 0.208305 |
| supply | 0.784235 | 0.179251 |
| and | 0.782442 | 0.177458 |
| highs | 0.392363 | 0.152981 |
| contraction | 0.161301 | 0.131145 |
| debt | 0.343350 | 0.103967 |



| | losses | difference |
|---|---|---|
| climbed | 0.060755 | -0.544229 |
| by | 0.074233 | -0.530751 |
| higher | 0.101645 | -0.503339 |
| above | 0.189008 | -0.415976 |
| $100 | 0.341984 | -0.262999 |
| geopolitical | 0.346327 | -0.258656 |
| barrel, | 0.397324 | -0.207660 |
| tensions | 0.404306 | -0.200678 |
| as | 0.048798 | -0.190585 |
| to | 0.459402 | -0.145582 |
| for | 0.480292 | -0.124692 |
| inflation | 0.123758 | -0.115624 |
| consumers' | 0.159512 | -0.079871 |
| rising | 0.162198 | -0.077185 |
| hit | 0.169523 | -0.069859 |
| driven | 0.550782 | -0.054202 |
| India. | 0.231174 | -0.051014 |
| U.S. | 0.231750 | -0.050438 |
| consumers. | 0.563851 | -0.041133 |
| have | 0.241770 | -0.040419 |

These are the token deletion test results for sentences originally predicted as negative. The left figure shows the top 20 highest values, while the right figure shows the top 20 lowest values.

The left table highlights the top 20 tokens that cause the largest increases in loss values when removed from sentences originally predicted as negative. These tokens play a critical role in maintaining the model's negative sentiment predictions. Tokens such as *"disruptions,"*, *"leading"*, *"prices,"*, and *"budgets."* show the highest loss differences, indicating their strong contribution to the negative sentiment classification. These words are often associated with financial or economic difficulties, aligning well with the negative sentiment context. Other domain-relevant tokens like *"affected,"*, *"costs,"*, *"businesses,"*, and *"dropped"* also significantly influence the model's predictions, reflecting their importance in shaping the overall sentiment. Even structural or quantitative terms like *"$20,000,"* and *"per"* contribute meaningfully, suggesting the model recognizes their relevance in the specific negative contexts.

The right table highlights the top 20 tokens with the smallest differences in loss values when removed from sentences originally predicted as negative. These tokens contribute minimally to the model's predictions and, in some cases, even negatively affect the accuracy of its original negative sentiment classification. Tokens such as *"climbed,"*, *"higher,"*, *"above,"*, and *"$100"* show the largest negative differences, indicating that their removal slightly reduces the loss. These tokens are typically associated with positive or neutral contexts, which might conflict with the overall negative sentiment of the sentences, thereby hindering the model's ability to make accurate predictions.

# 5. Summary

This report evaluates the interpretability of a sentiment analysis model using three methods: attention analysis, sentiment score perturbation, and loss-based perturbation. Key tokens like *"good,"*, *"soaring,"*, *"opportunities,"*, *"disruptions,"*, and *"budgets."* consistently emerged as critical for predictions, reflecting the model's reliance on sentiment-laden and domain-specific words. Structural words like *"as"* and *"Despite"* also played contextual roles, while some neutral or positive tokens, such as *"climbed"* and *"$100"*, negatively influenced predictions in negative contexts, highlighting areas for improvement. The findings underscore the model's strengths in leveraging key tokens but reveal vulnerabilities in handling ambiguous inputs.