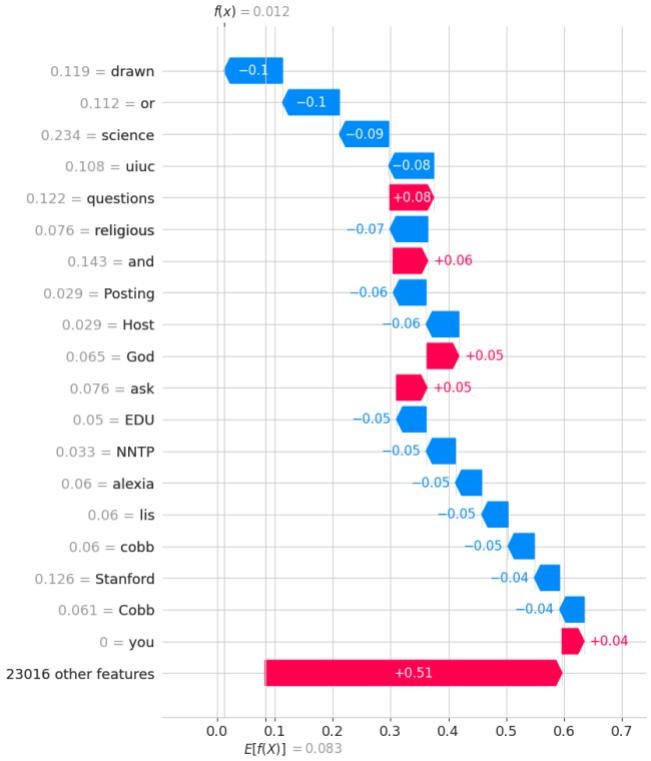
Q1

In the lecture titled "Content Moderation in Social Media and AI," the data science issues discussed have implications for a variety of stakeholders, including social media platforms, consumers, moderators, content creators, and society. One potential advantage of data science applications in content moderation is that they may aid in preventing the spread of harmful or illegal content, thereby safeguarding vulnerable populations and fostering a safer online environment. There are, however, potential negative effects, such as the unintended removal of harmless content and the suppression of free expression.

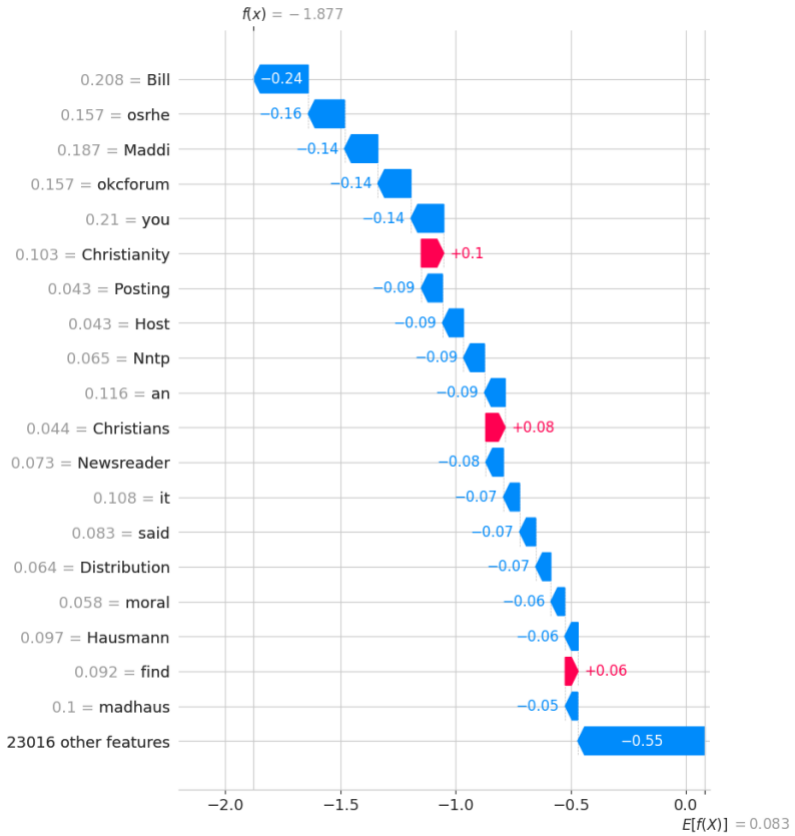
The lecture addressed the proprietorship of data and the extent to which privacy is protected as a concern pertaining to data protection and sharing. It is unclear how social media platforms and content moderators use, share, or safeguard the vast amounts of user data they frequently have access to. In addition, the lecture emphasizes the difficulties associated with balancing privacy concerns with the need for transparency and accountability in content moderation processes.

Transparency and interpretability are crucial for ensuring the fairness and impartiality of content moderation processes. As the lecture notes, however, many of the algorithms used for content moderation are "black boxes," meaning that their inner workings are not completely understood by users or transparent to other stakeholders. This lack of transparency makes it difficult to hold moderators or platforms accountable for their decisions and may contribute to discrimination and bias in content moderation.
The vendor's incentives have a significant impact on their data protection, transparency, and impartiality practices. In the case of social media platforms, their primary incentive is to maximize user engagement and profitability, which may conflict with the need for content moderation that is transparent and equitable. This misalignment of incentives could contribute to a lack of accountability in content moderation processes and erode user confidence.
The lecture concludes by highlighting the complex ethical challenges associated with content moderation in social media and artificial intelligence. To ensure that data science applications in content moderation are used responsibly and ethically, it is necessary to consider the interests and perspectives of all stakeholders, including consumers, content creators, moderators, and society as a whole. This necessitates a dedication to transparency, fairness, and accountability, as well as a contemplation of the potential unintended consequences of data science applications in content moderation.

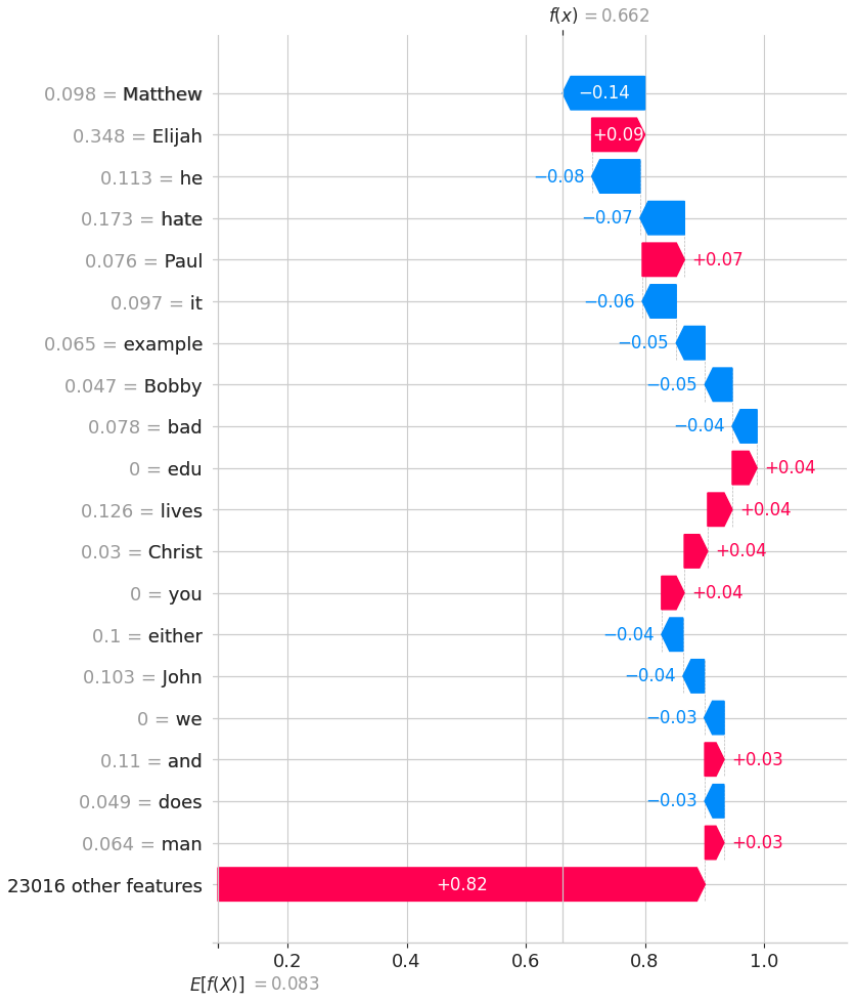
Q2
Part a: see colab for code
Part b
The confusion matrix is [284 35]
[3 395]
Explainer:
```
index: 4, the accuracy of prediction: False
```



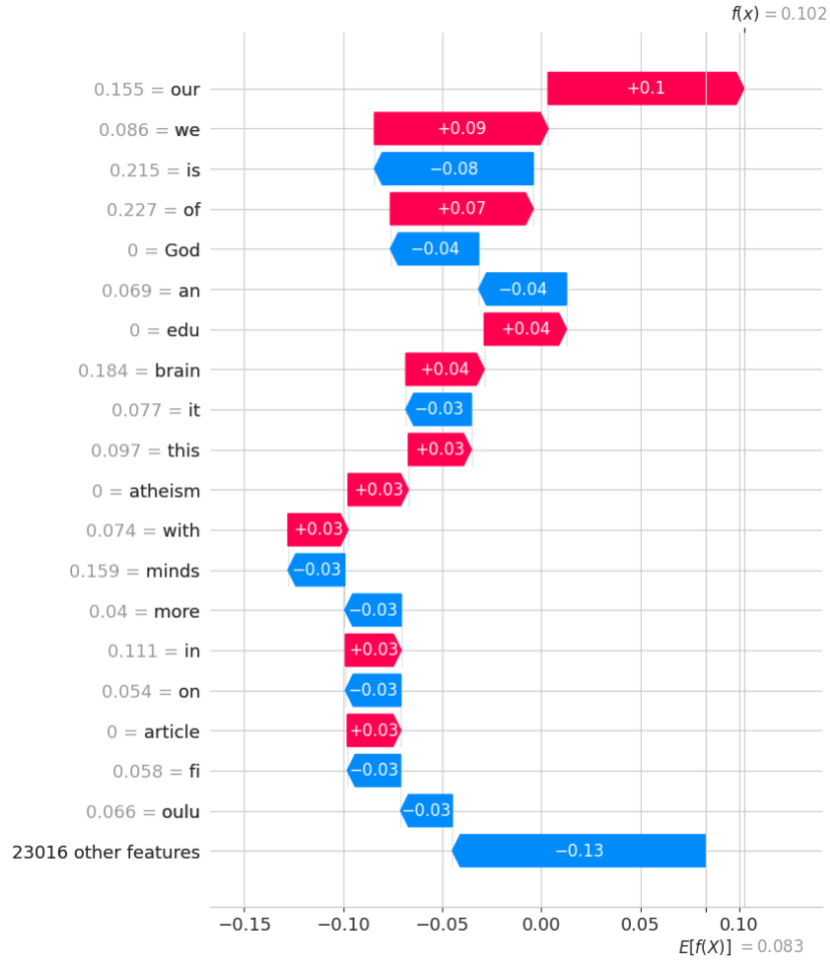
index: 185, the accuracy of prediction: True



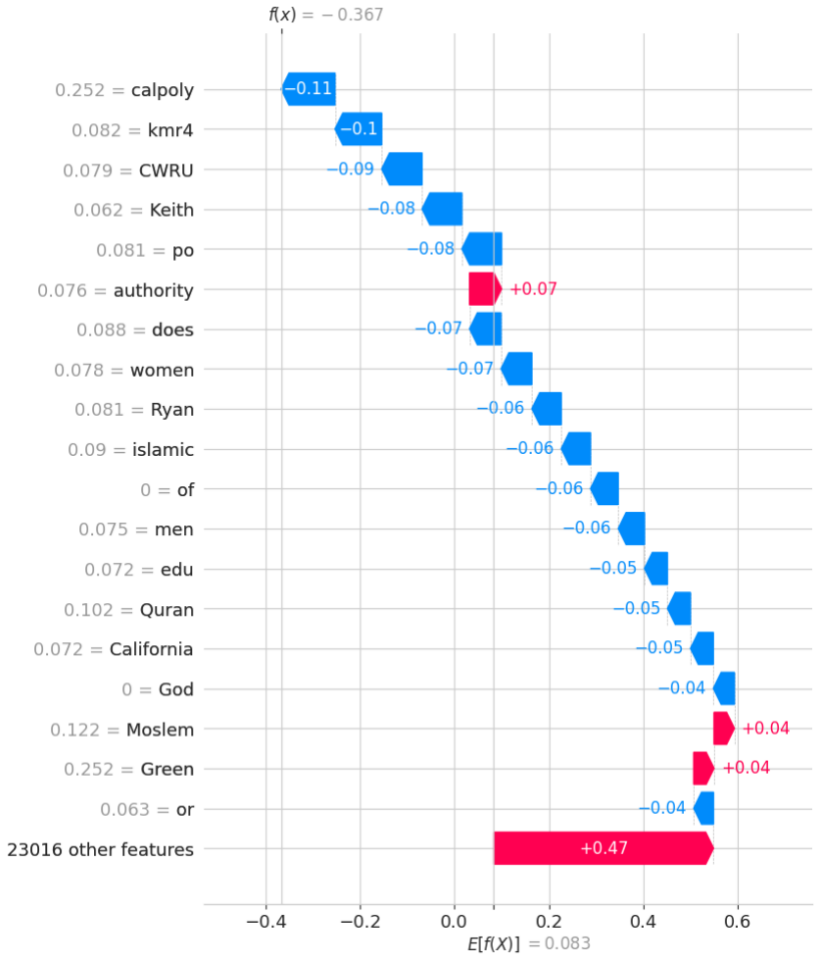
index: 347, the accuracy of prediction: True



index: 216, the accuracy of prediction: False

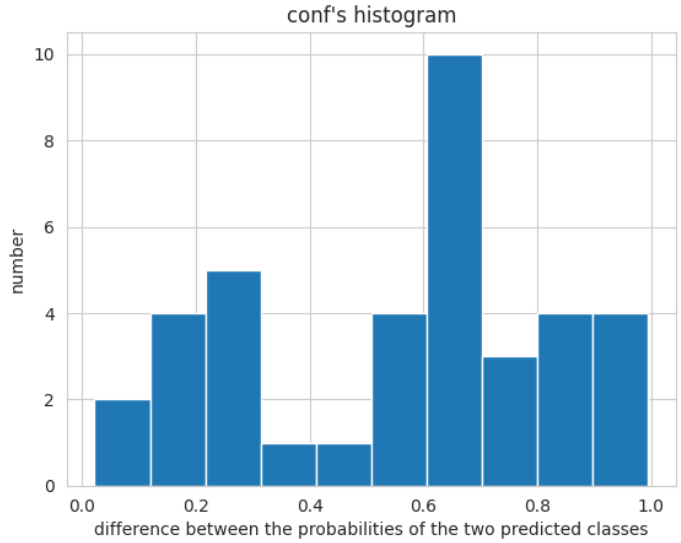


index: 223, the accuracy of prediction: True

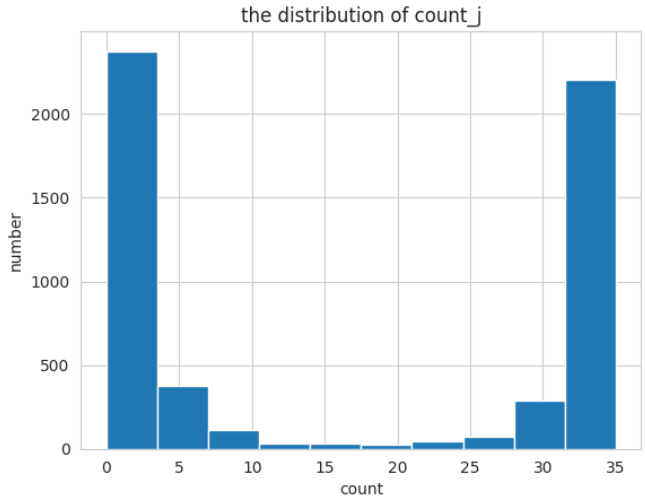


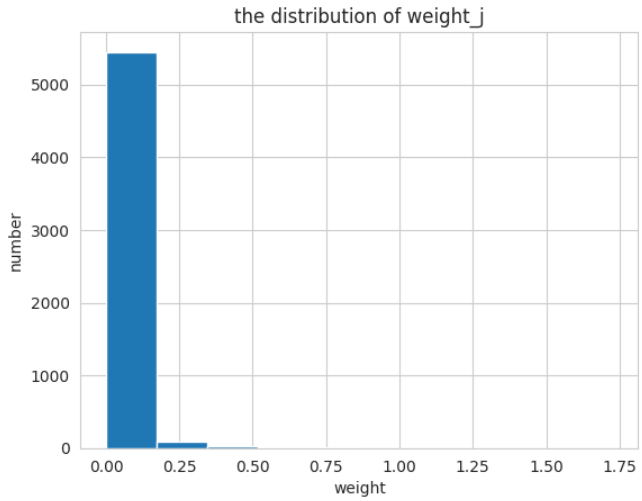
part c

1. the accuracy of the classifier: 0.9470013947001394
   the number of misclassified documents: 38
2. `Text(0.5, 1.0, "conf's histogram")`



Part 3





Count_j is somewhat bimodal with one mode at around 0 and the other at 35, whilst showing that count j is above 2000 at both of these nodes. There is a higher chance that count_j occurs at extreme values.

Weight j is a unimodal histogram showing over 5000 texts has a 0 weight and a little above 0(presumably 100) texts that has around 0.25 in weight.

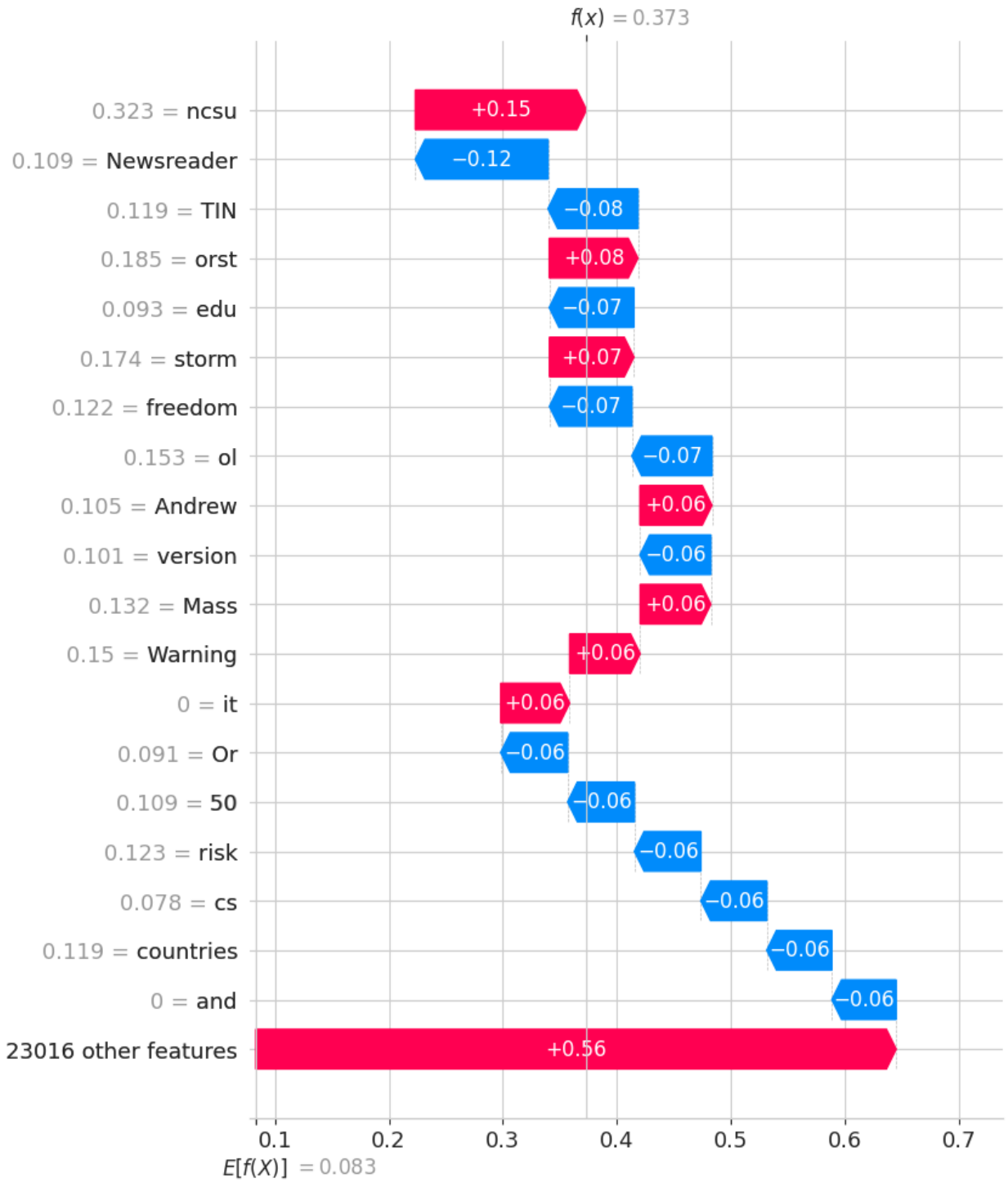
Judging from the distribution of count_j, the count_j of most words is distributed at both ends, indicating that the number of articles with words that have a negative impact is either large or very few. Follow-up adjustments should be made to words with a larger count_j. From the distribution of weight_j, The sum of the absolute values of the shap values of most negatively affected words is small, and the larger ones only account for a small part, and the larger part can be modified later

Part d

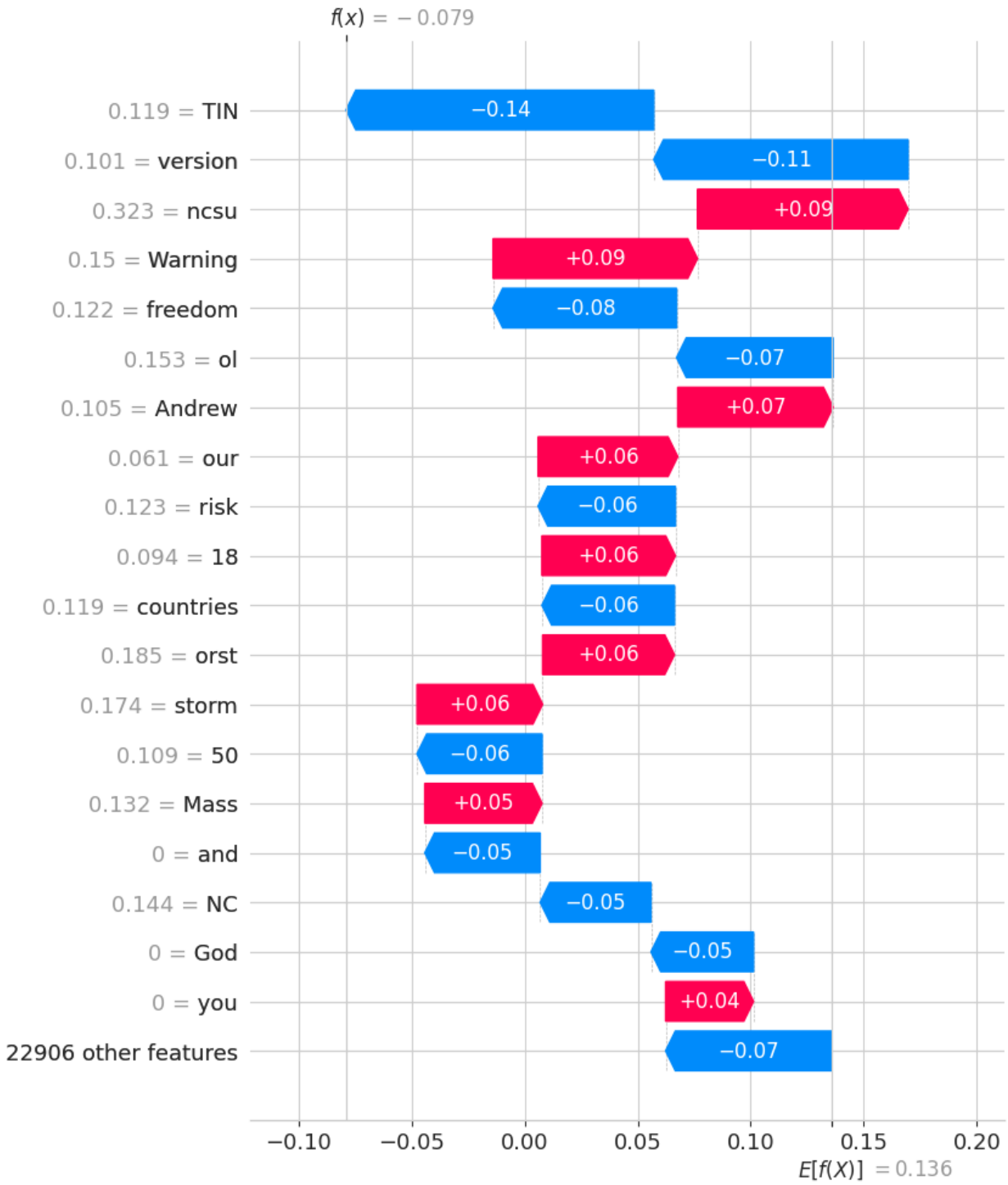
the accuracy of the classifier: 0.9539748953974896

the number of misclassified documents: 33

graph before:


f(x) = 0.373
0.323 = ncsu
0.109 = Newsreader
0.119 = TIN
0.185 = orst
0.093 = edu
0.174 = storm
0.122 = freedom
0.153 = ol
0.105 = Andrew
0.101 = version
0.132 = Mass
0.15 = Warning
0 = it
0.091 = Or
0.109 = 50
0.123 = risk
0.078 = cs
0.119 = countries
0 = and
23016 other features
+0.15
−0.12
−0.08
+0.08
−0.07
+0.07
−0.07
−0.07
+0.06
−0.06
+0.06
+0.06
+0.06
−0.06
−0.06
−0.06
−0.06
−0.06
−0.06
+0.56
0.1
0.2
0.3
0.4
0.5
0.6
0.7
E[f(X)] = 0.083

Graph after:



In the first example you can see newsreader dragging down the predicted value. In the second example, while the expected value rose by 0.05, the predicted value dropped. The reason behind that could be due to the Tin variable lowering the predicted value by 0.14.

Delete words with count>20 and weight>=0.12, and then SGDClassifier conducts training and prediction, and finds that the number of misclassified documents has become 33, which is 5 fewer than the previous 38, so it can be found to delete these words is positive, At the same

time, we selected the fourth sample, which was misclassified before, but now it is correctly classified. It can be found that the modified shape values are much more reasonable than the previous ones.

It can be seen that the features that play a negative role do not play such a large negative role in the new model, so they can be correctly classified