SAEfarer User Study

Introduction, tutorial, and interview slides

Introduction

- We have developed an interactive visualization tool called SAEfarer for exploring concepts learned by text classification models
- SAEfarer runs in Jupyter notebooks
- We are evaluating this tool with machine learning researchers

Agenda

- Sparse autoencoder overview (~5 min.)
- SAEfarer tutorial (~20 min.)
- SAEfarer practice (~20 min.)
- Model exploration (~30 min.)
- Questionnaire and interview (~15 min.)

Sparse Autoencoders

Sparse autoencoders (SAEs)

- SAEs are a technique for interpreting transformer-based language models
- It is difficult to interpret an LM’s neuron activations
- Train an SAE to project the neuron activations at a chosen layer in the LM to a sparse, higher-dimensional space that is easier to interpret
- Each dimension of this new space ideally represents a single, human-understandable concept
- These dimensions or concepts are referred to as **features** of the SAE

Sparse autoencoders (SAEs)

- The goal of the SAE is to identify interpretable features used by the model
- When an instance is input to the model, every token will activate a small number of the SAE features
- The activation value of an SAE feature represents the strength of the presence of the given concept in that token
    - If a feature’s activation value is 0 for a given token, then the concept is not present in that token

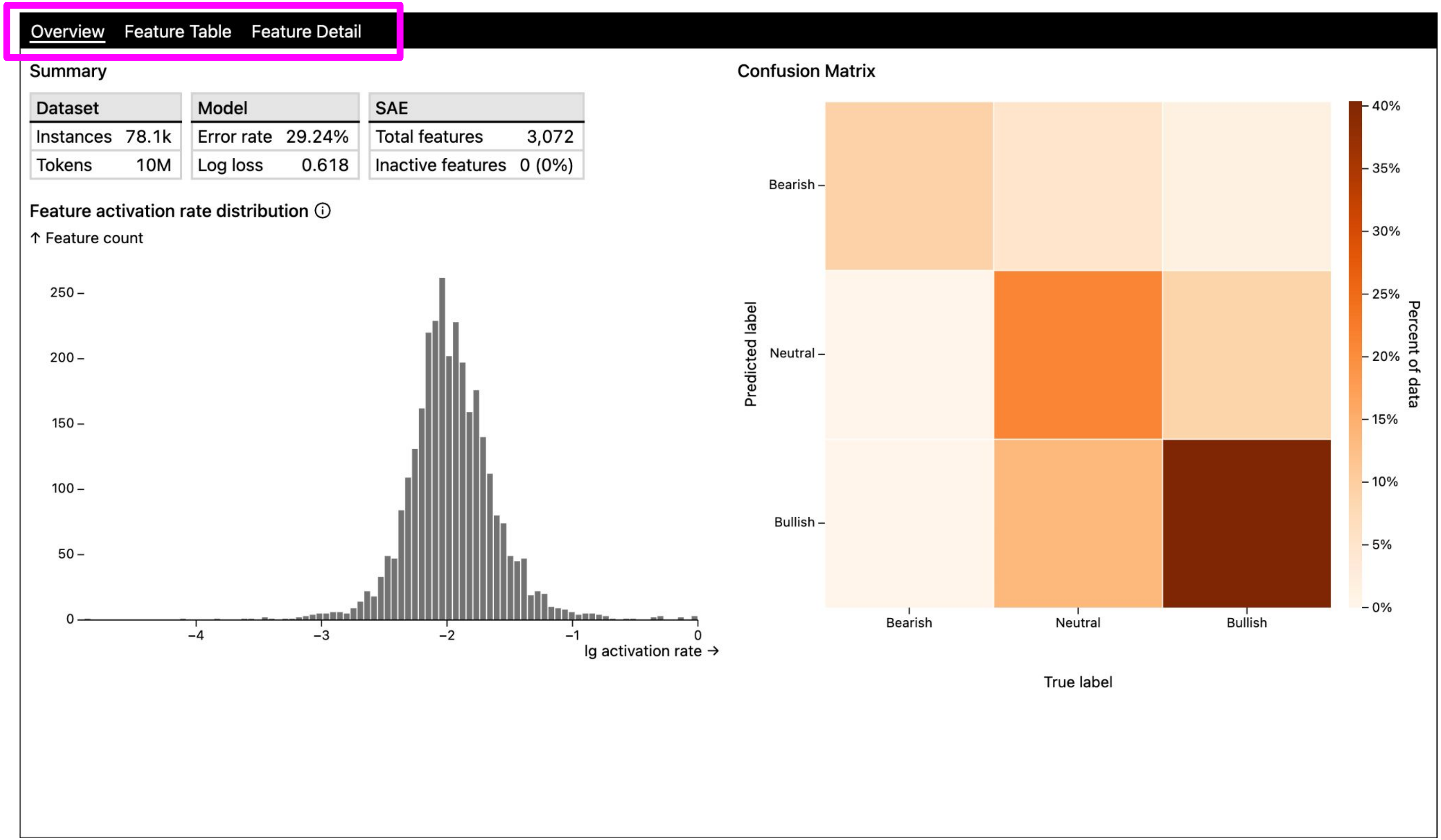
Sparse autoencoders (SAEs)

- The goal of the SAE is to identify interpretable features used by the model
- Text classification:
  - Each data point or instance is a piece of text that is represented by a series of tokens
  - Each instance has a ground truth label
  - When the model makes a prediction, it assigns a probability to each class
- We are interested in exploring the relationship between the SAE features and the model’s predictions
  - What instances cause a given feature to activate?
  - Are there any trends in the model’s predictions on those instances?

SAEfarer Tutorial

SAEfarer

- Interactive visualization tool that runs in Jupyter notebook
- Analyzing text classification models
- Exploring the relationships between the SAE features and the model’s predictions and errors
- Example
    - cryptobert: RoBERTa-based model that classifies the sentiment of cryptocurrency social media posts
    - Three classes:
        - Bearish (negative)
        - Neutral
        - Bullish (positive)



Summary

Dataset	
Instances	78.1k
Tokens	10M

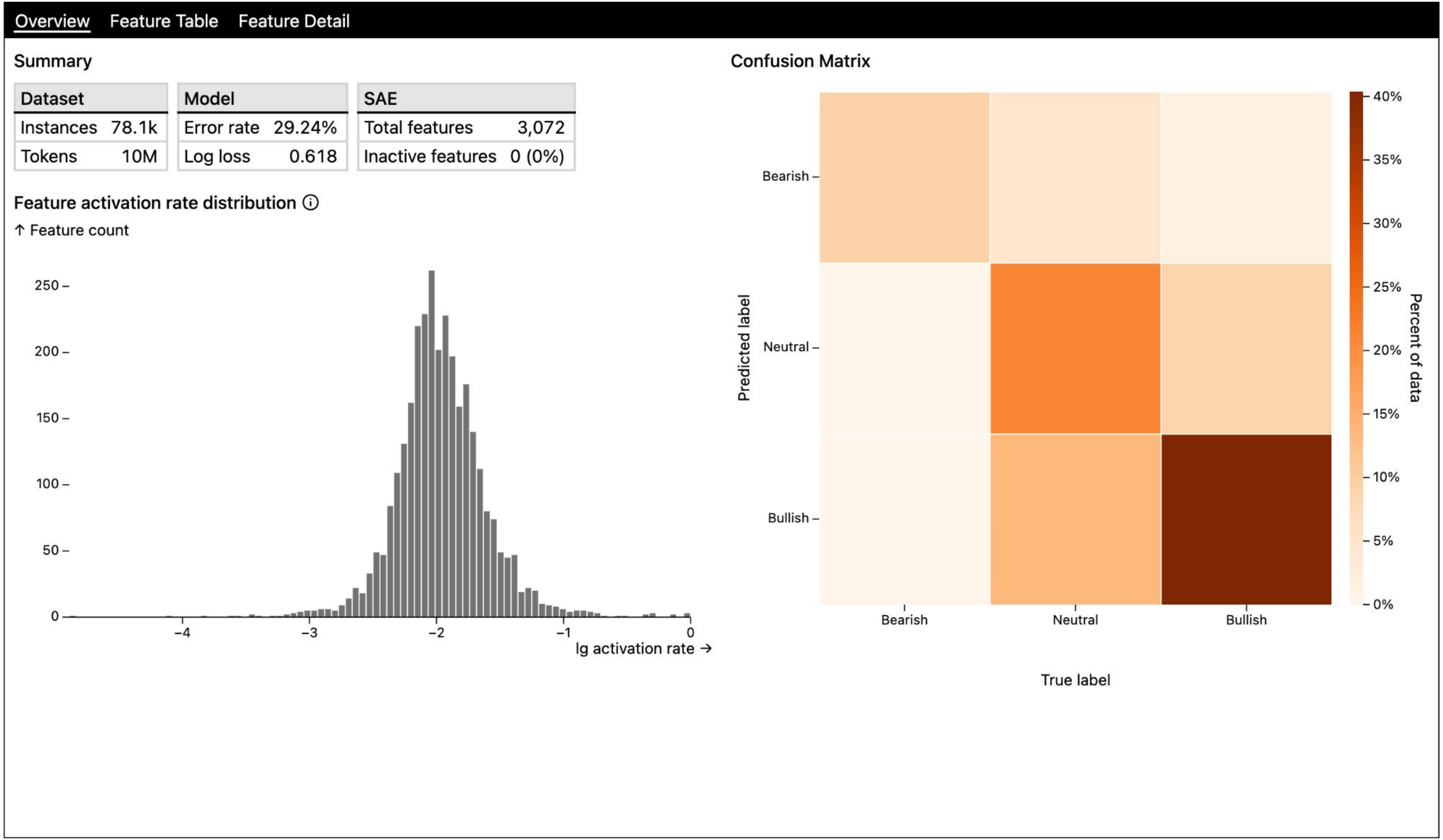
Model	
Error rate	29.24%
Log loss	0.618

SAE	
Total features	3,072
Inactive features	0 (0%)

Feature activation rate distribution ⓘ

Confusion Matrix

The UI has three tabs at the top.



Summary

Dataset	
Instances	78.1k
Tokens	10M

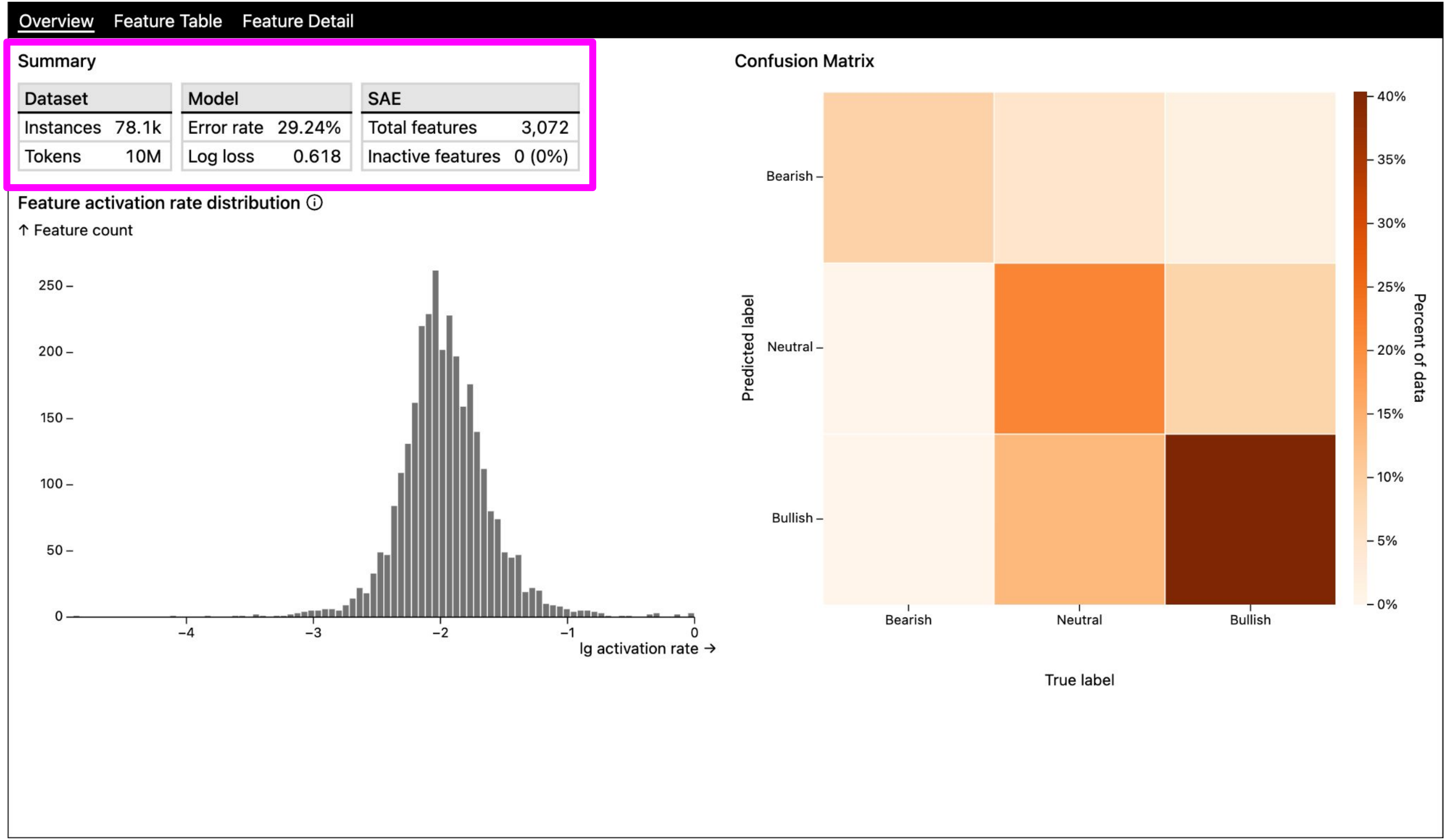
Model	
Error rate	29.24%
Log loss	0.618

SAE	
Total features	3,072
Inactive features	0 (0%)

Feature activation rate distribution ⓘ

Confusion Matrix

First, we'll cover the Overview tab.



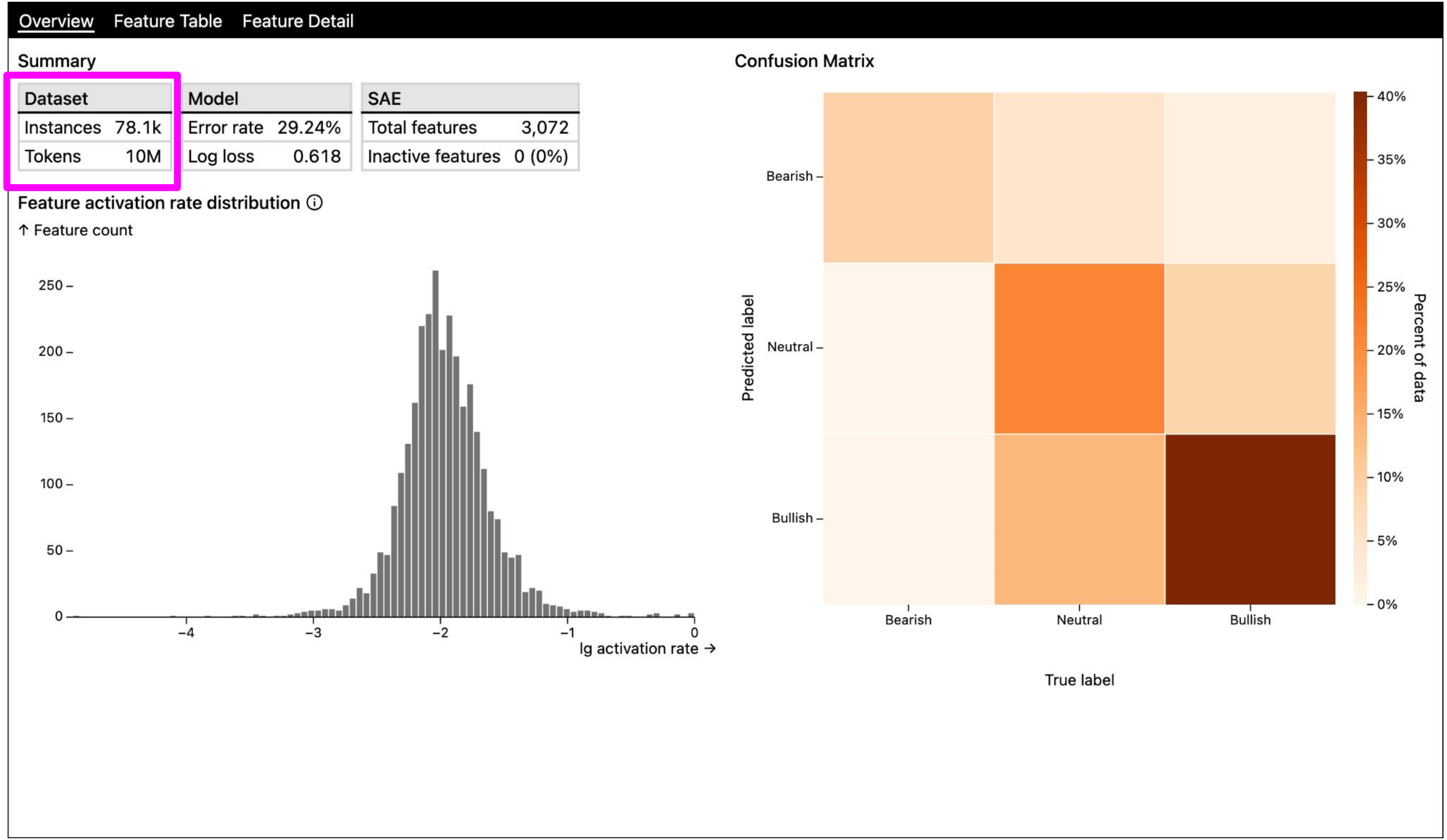
Summary

Dataset	
Instances	78.1k
Tokens	10M

Model	
Error rate	29.24%
Log loss	0.618

SAE	
Total features	3,072
Inactive features	0 (0%)

At the top, there is summary information about the dataset, model and SAE.



Summary

Dataset	
Instances	78.1k
Tokens	10M

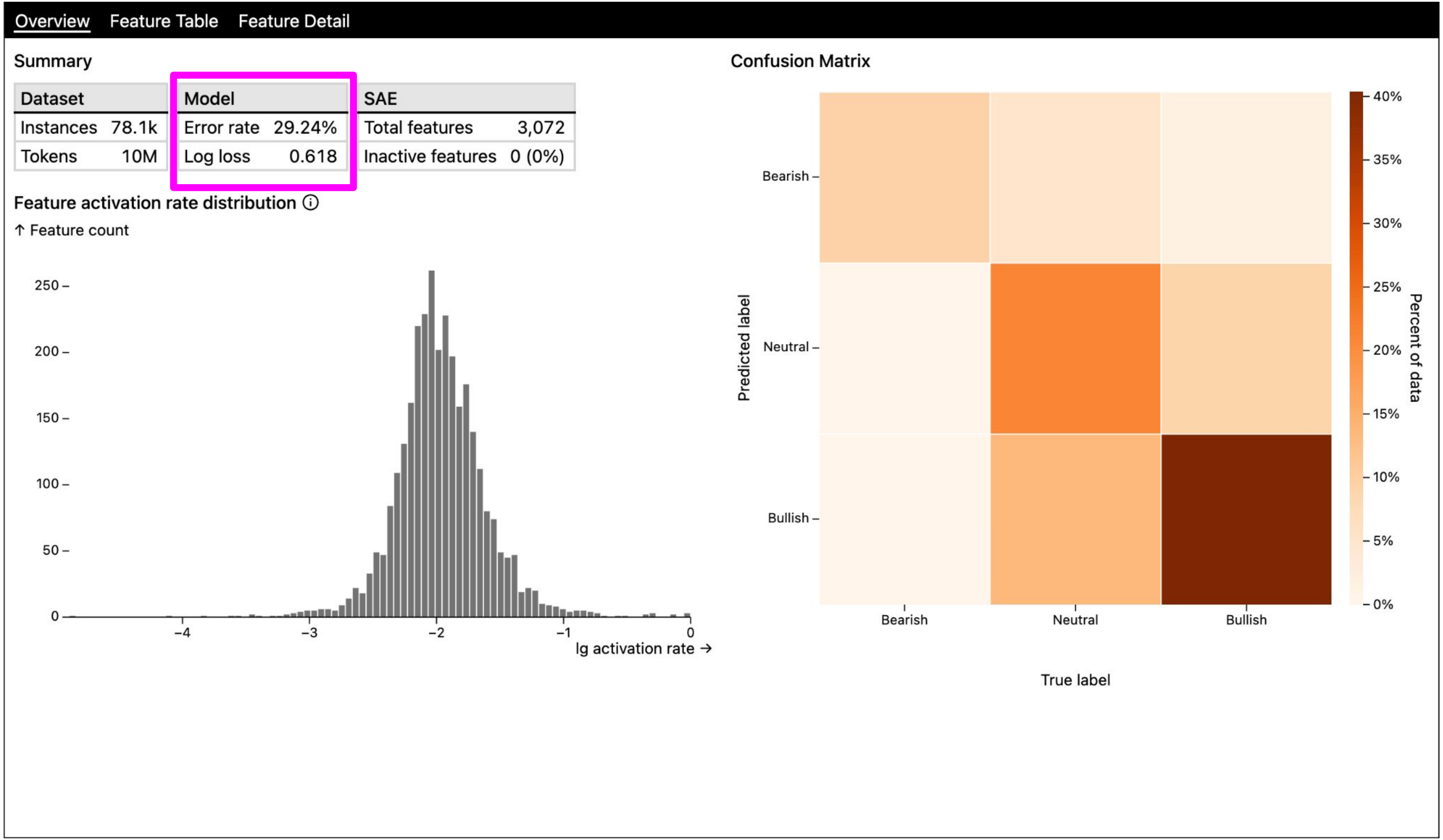
Model	
Error rate	29.24%
Log loss	0.618

SAE	
Total features	3,072
Inactive features	0 (0%)

Feature activation rate distribution ⓘ

Confusion Matrix

78,000 instances (10 million tokens) were used to analyze the SAE.



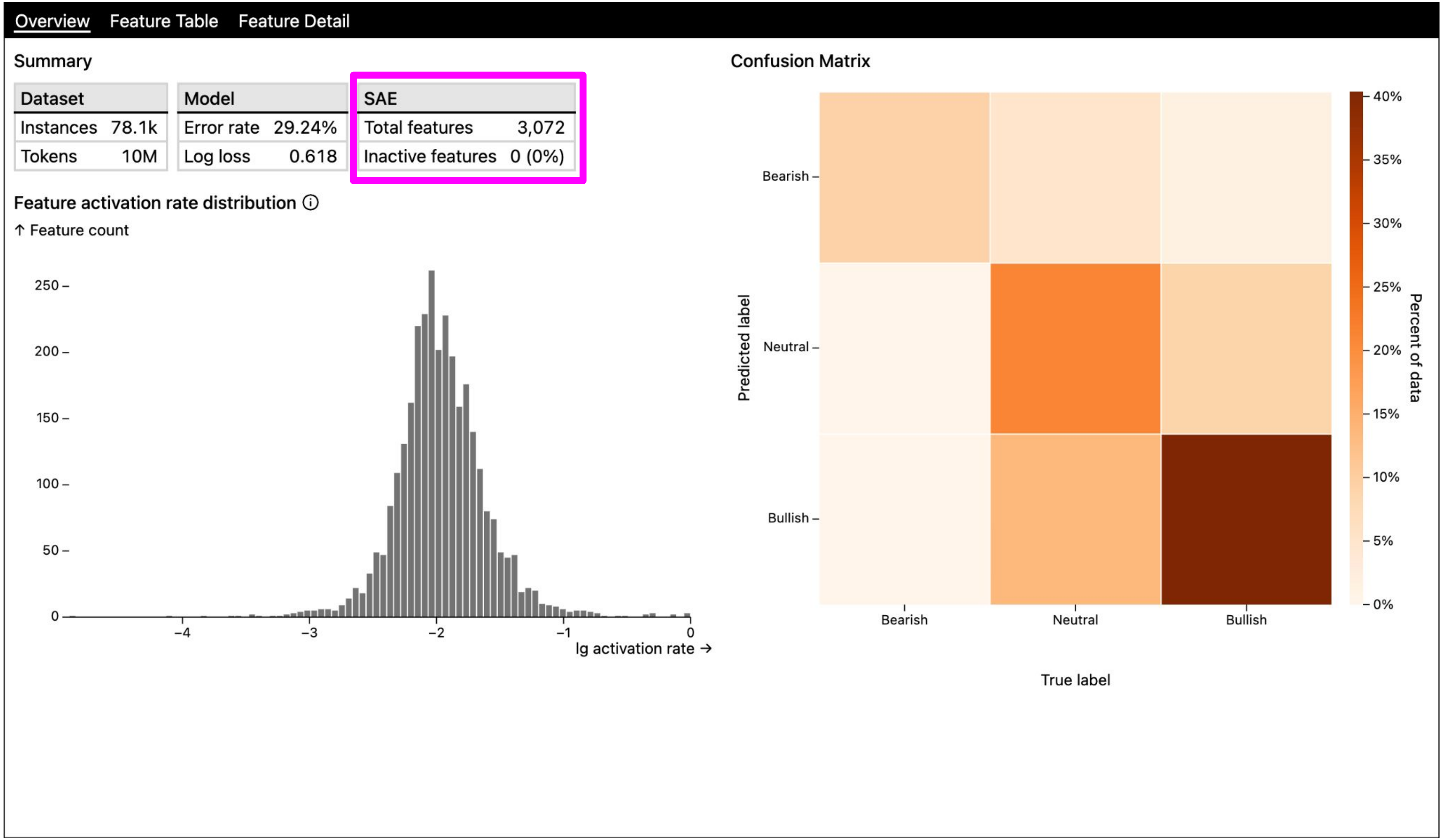
Summary

Dataset	
Instances	78.1k
Tokens	10M

Model	
Error rate	29.24%
Log loss	0.618

SAE	
Total features	3,072
Inactive features	0 (0%)

The error rate and log loss of the classification model on these instances.



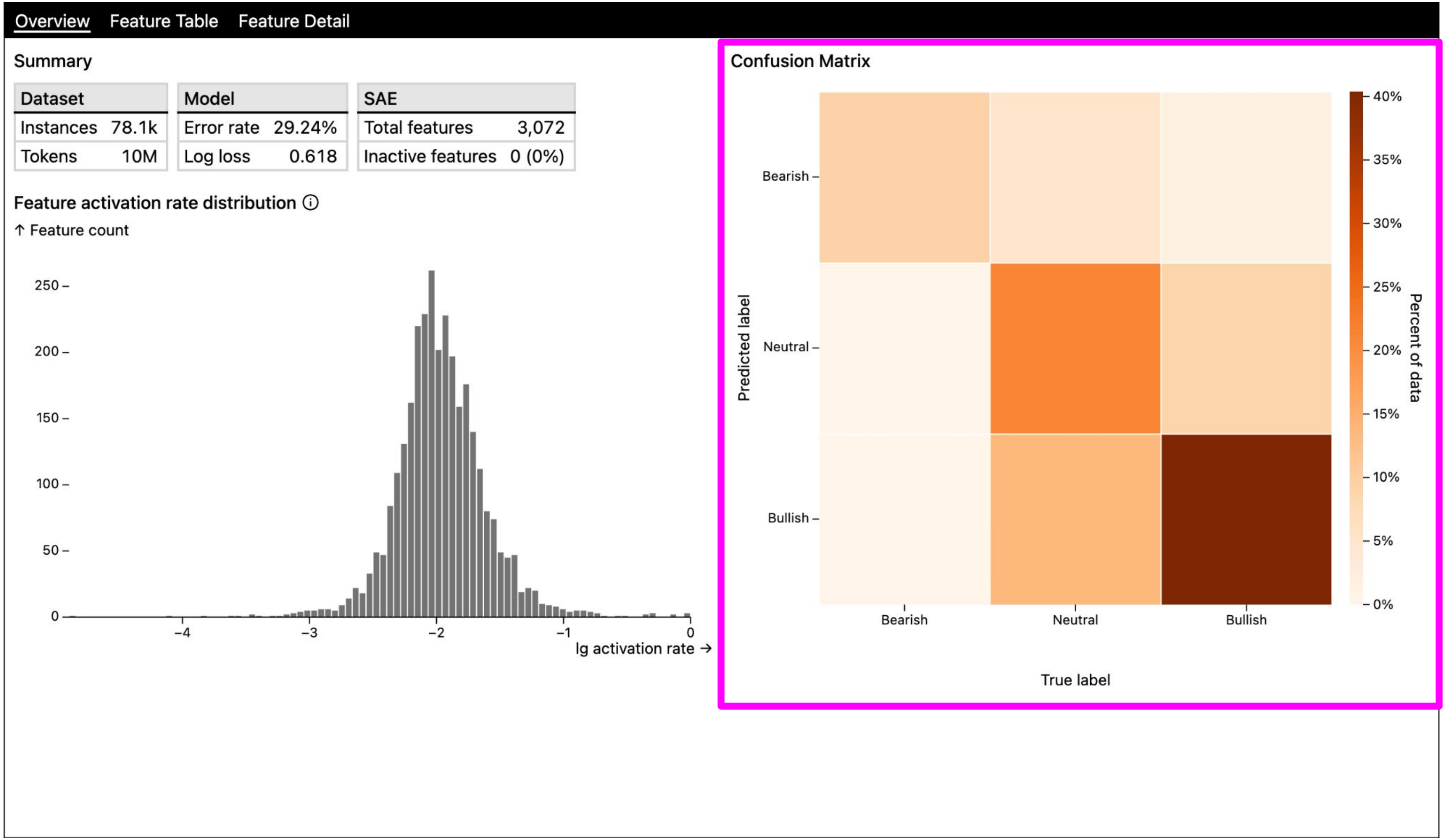
Summary

Dataset	
Instances	78.1k
Tokens	10M

Model	
Error rate	29.24%
Log loss	0.618

SAE	
Total features	3,072
Inactive features	0 (0%)

The SAE has 3072 features. All of them activated during analysis.



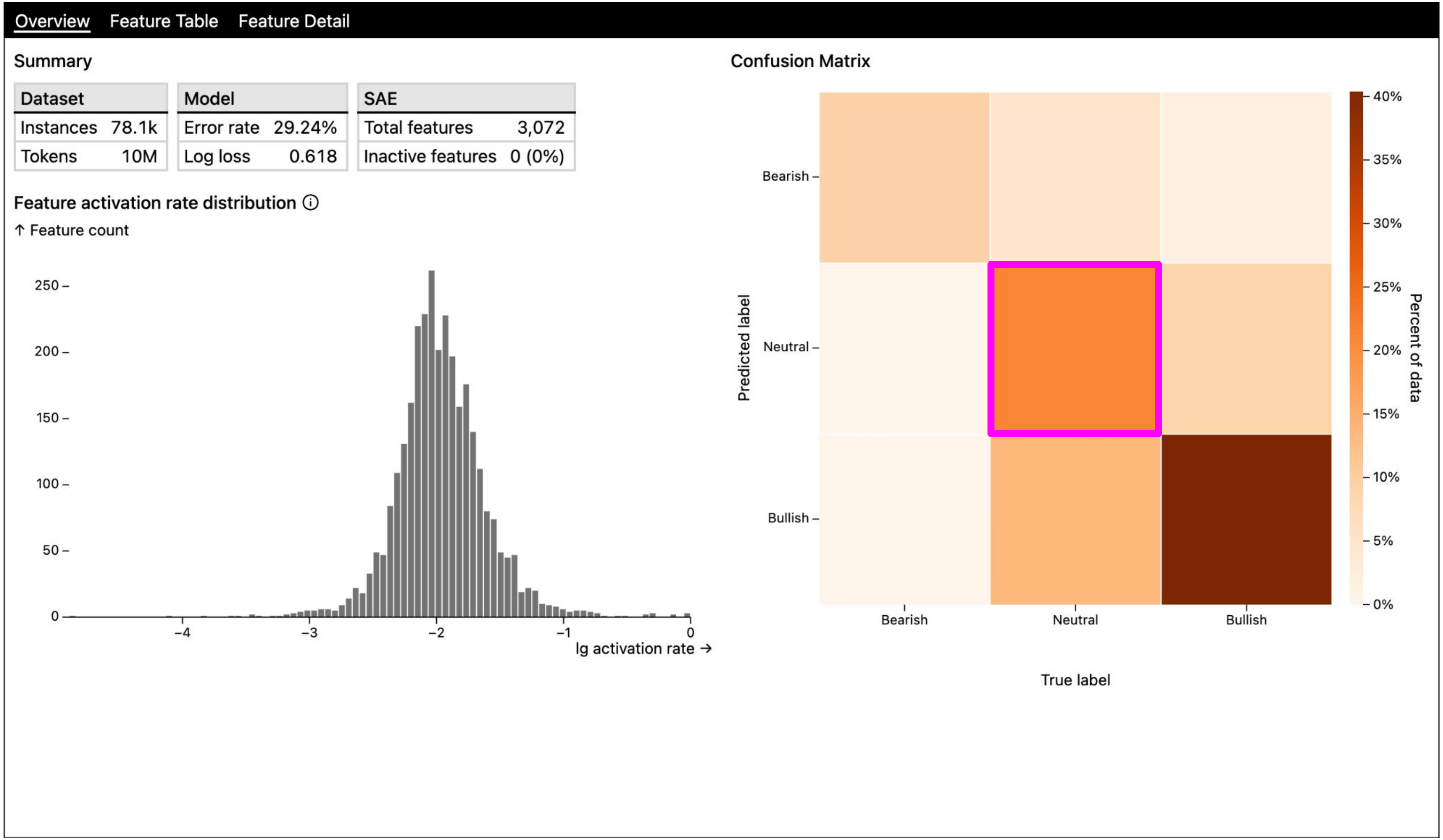
Summary

Dataset	
Instances	78.1k
Tokens	10M

Model	
Error rate	29.24%
Log loss	0.618

SAE	
Total features	3,072
Inactive features	0 (0%)

The confusion matrix shows the model’s performance.



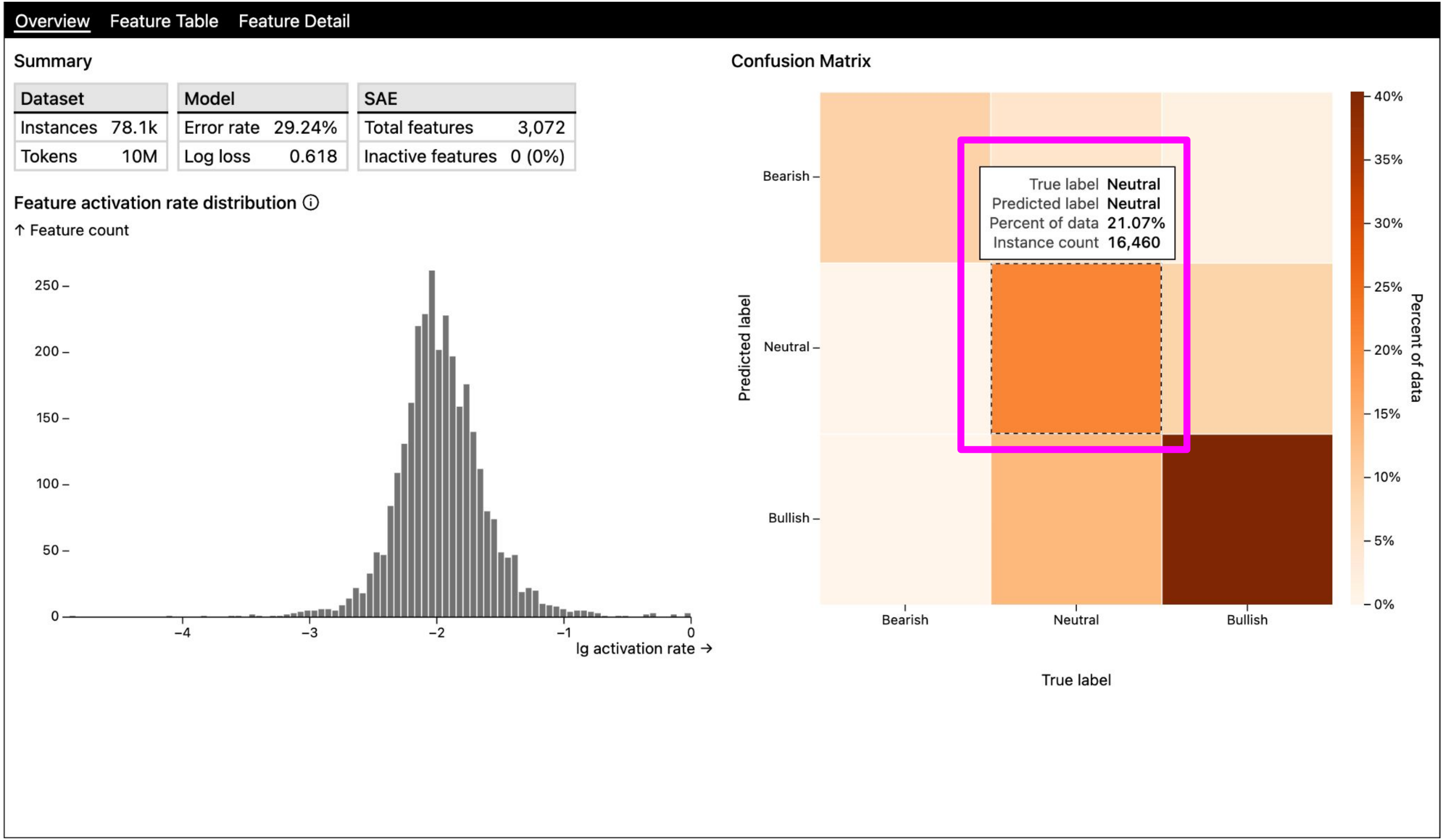
Summary

Dataset	
Instances	78.1k
Tokens	10M

Model	
Error rate	29.24%
Log loss	0.618

SAE	
Total features	3,072
Inactive features	0 (0%)

Each cell shows the percentage of instances with the given predicted and true label.



Summary

Dataset	
Instances	78.1k
Tokens	10M

Model	
Error rate	29.24%
Log loss	0.618

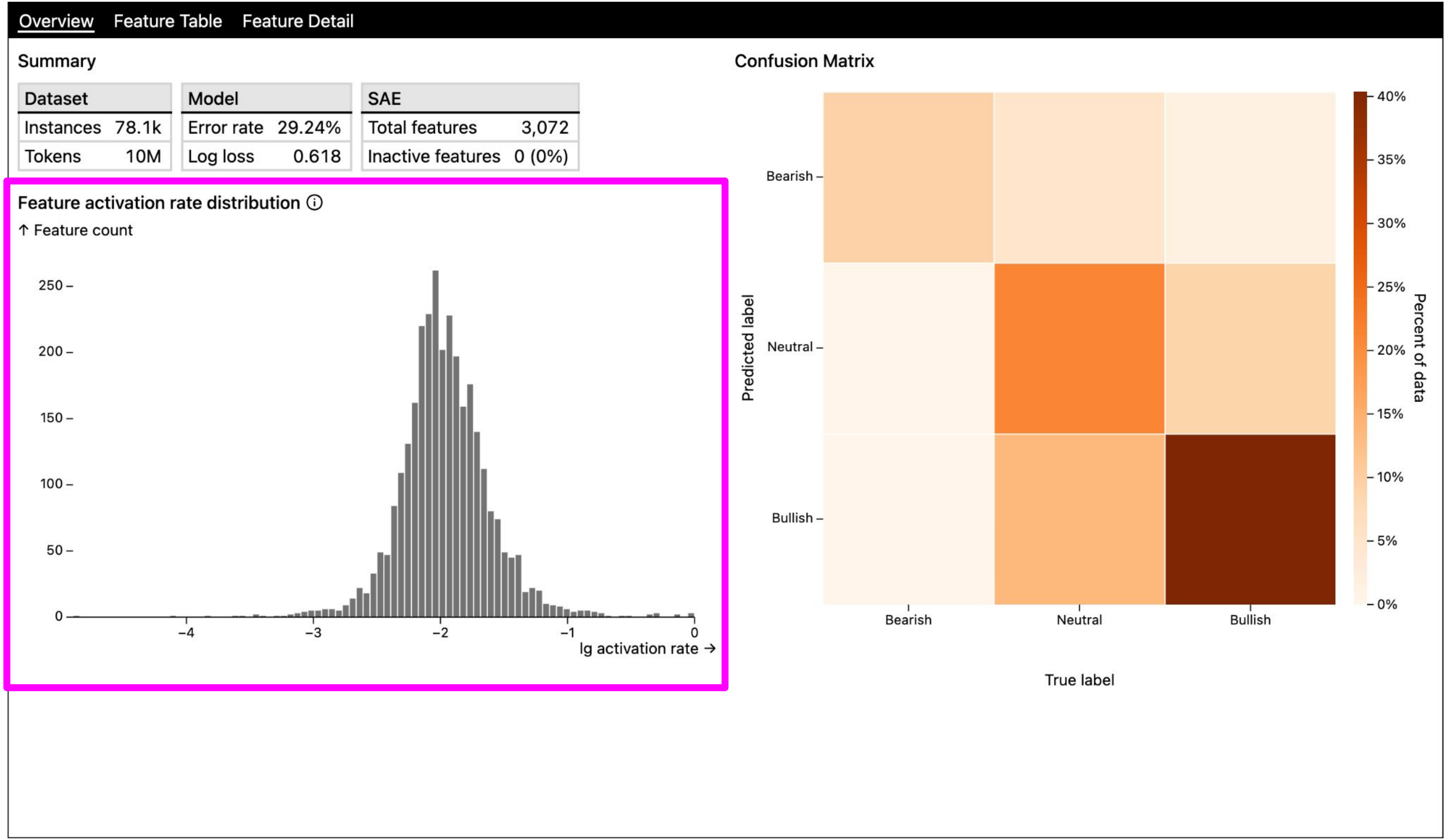
SAE	
Total features	3,072
Inactive features	0 (0%)

Feature activation rate distribution ⓘ

Confusion Matrix

True label	Neutral
Predicted label	Neutral
Percent of data	21.07%
Instance count	16,460

Most visualizations in SAEfarer provide details when you hover over them.



Summary

Dataset	
Instances	78.1k
Tokens	10M

Model	
Error rate	29.24%
Log loss	0.618

SAE	
Total features	3,072
Inactive features	0 (0%)

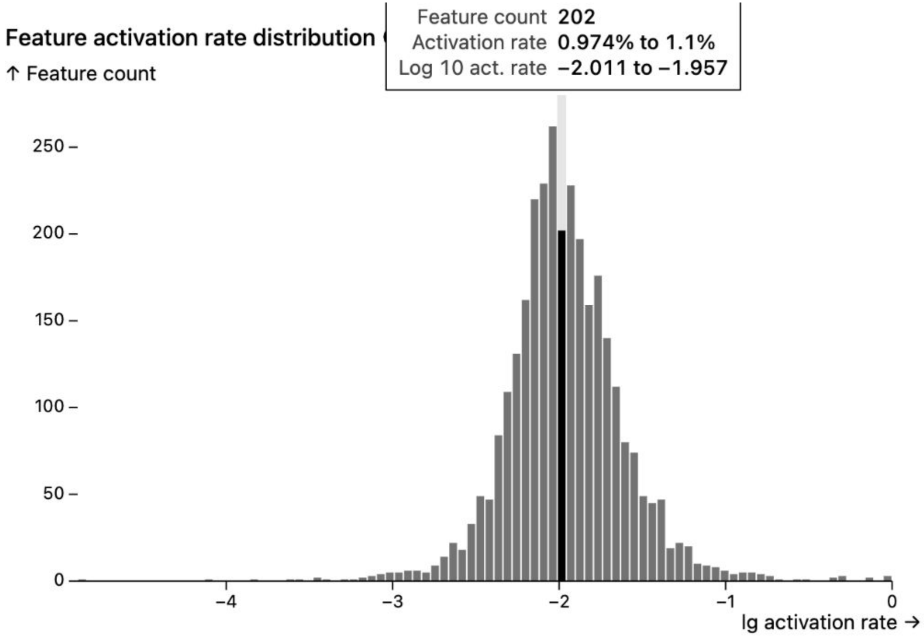
Feature activation rate distribution ⓘ

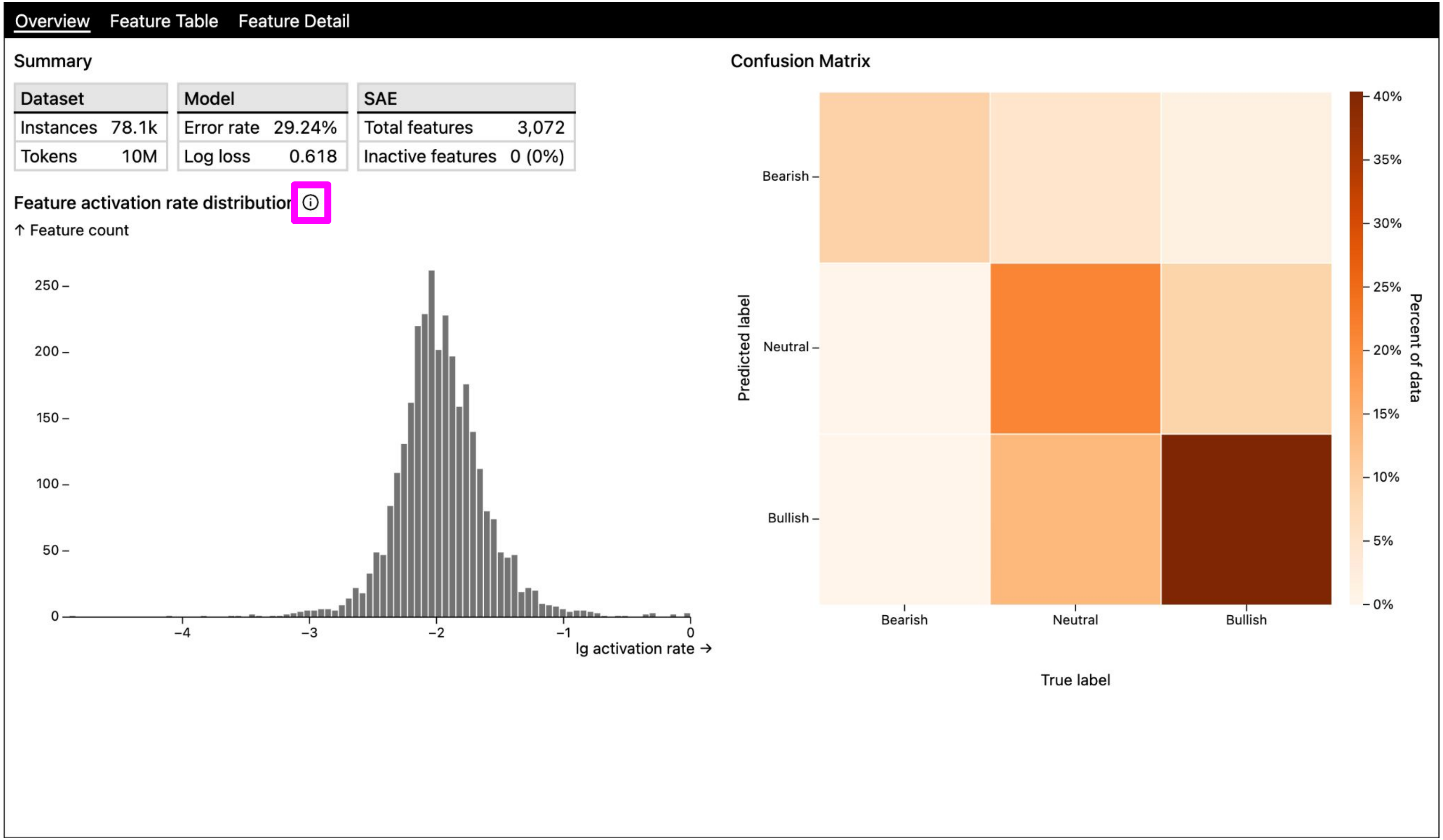
Confusion Matrix

The histogram shows the distribution of how often the SAE features activated.

Feature activation rate distribution

- Activation rate: The percentage of instances that cause the SAE feature to activate.
- This histogram uses a log10 scale on the x-axis.
  - ○ 10% act. rate = -1 lg act. rate
  - ○ 1% act. rate = -2 lg act. rate
- Example: 202 features activate on 0.974% to 1.1% of instances.





Summary

Dataset	
Instances	78.1k
Tokens	10M

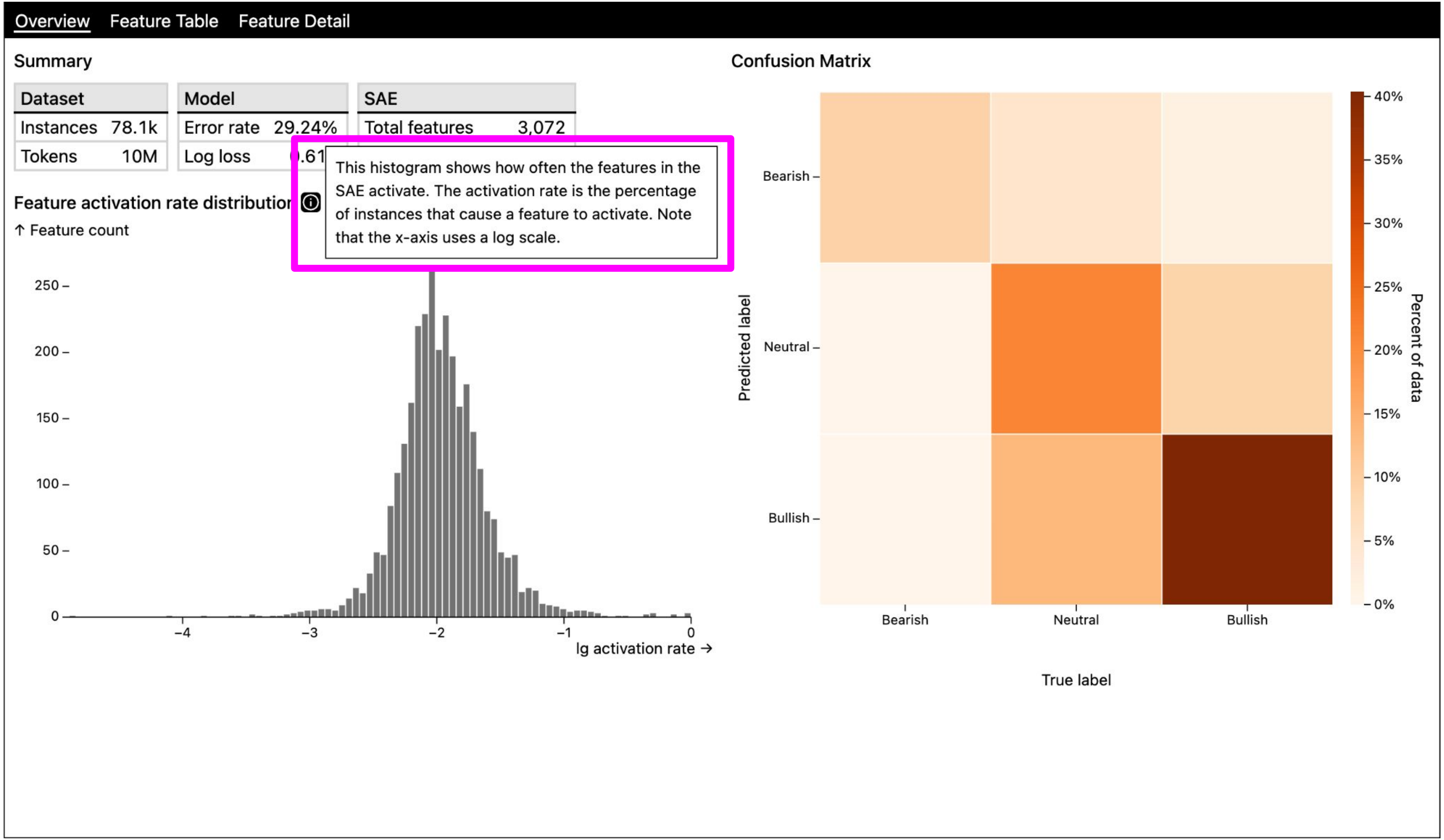
Model	
Error rate	29.24%
Log loss	0.618

SAE	
Total features	3,072
Inactive features	0 (0%)

Feature activation rate distribution ⓘ

Confusion Matrix

Throughout the UI, there are help icons.



Summary

Dataset	
Instances	78.1k
Tokens	10M

Model	
Error rate	29.24%
Log loss	0.61

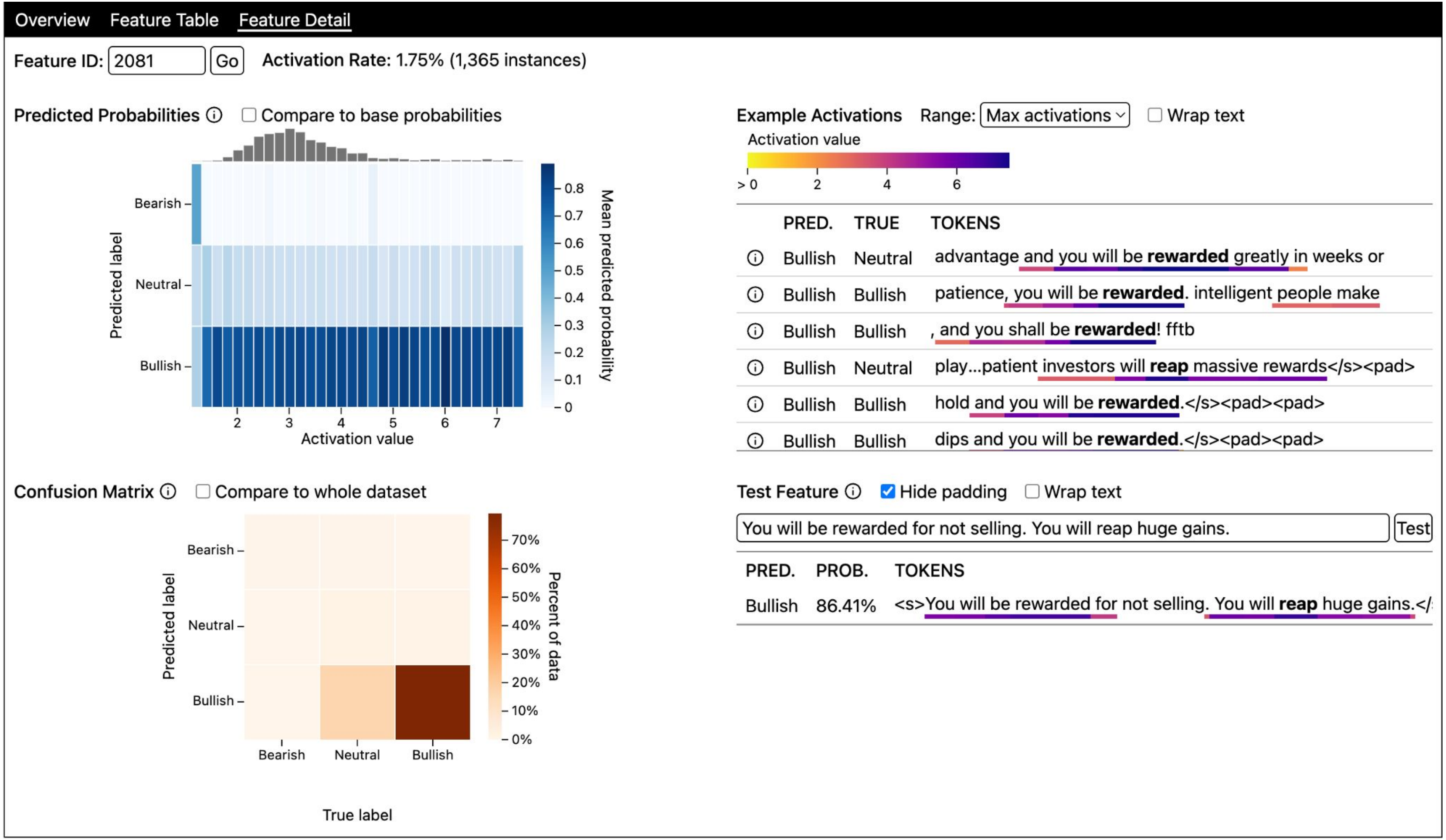
SAE	
Total features	3,072

Feature activation rate distribution

This histogram shows how often the features in the SAE activate. The activation rate is the percentage of instances that cause a feature to activate. Note that the x-axis uses a log scale.

Confusion Matrix

Hovering over the icon shows a tooltip that explains that part of the interface.



Feature ID: 2081

Activation Rate: 1.75% (1,365 instances)

Range: Max activations

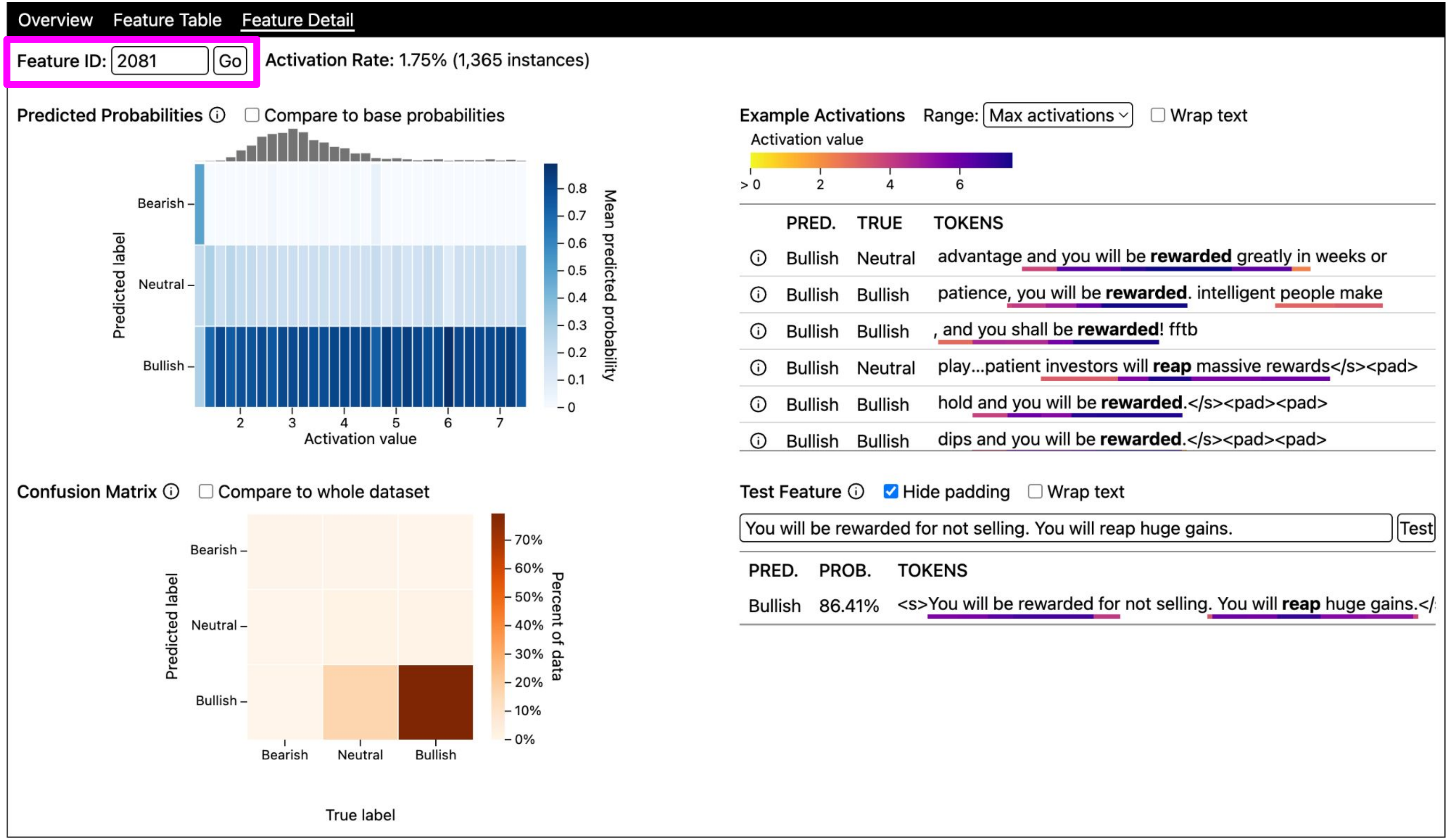
	PRED.	TRUE	TOKENS
ⓘ	Bullish	Neutral	advantage and you will be **rewarded** greatly in weeks or
ⓘ	Bullish	Bullish	patience, you will be **rewarded**. intelligent people make
ⓘ	Bullish	Bullish	, and you shall be **rewarded**! fftb
ⓘ	Bullish	Neutral	play...patient investors will **reap** massive rewards</s><pad>
ⓘ	Bullish	Bullish	hold and you will be **rewarded**.</s><pad><pad>
ⓘ	Bullish	Bullish	dips and you will be **rewarded**.</s><pad><pad>

☑ Hide padding □ Wrap text

You will be rewarded for not selling. You will reap huge gains.

PRED.	PROB.	TOKENS
Bullish	86.41%	<s>You will be rewarded for not selling. You will **reap** huge gains.</

Next, we'll cover the Feature Detail tab.



Feature ID: 2081 Go

Activation Rate: 1.75% (1,365 instances)

Range: Max activations □ Wrap text

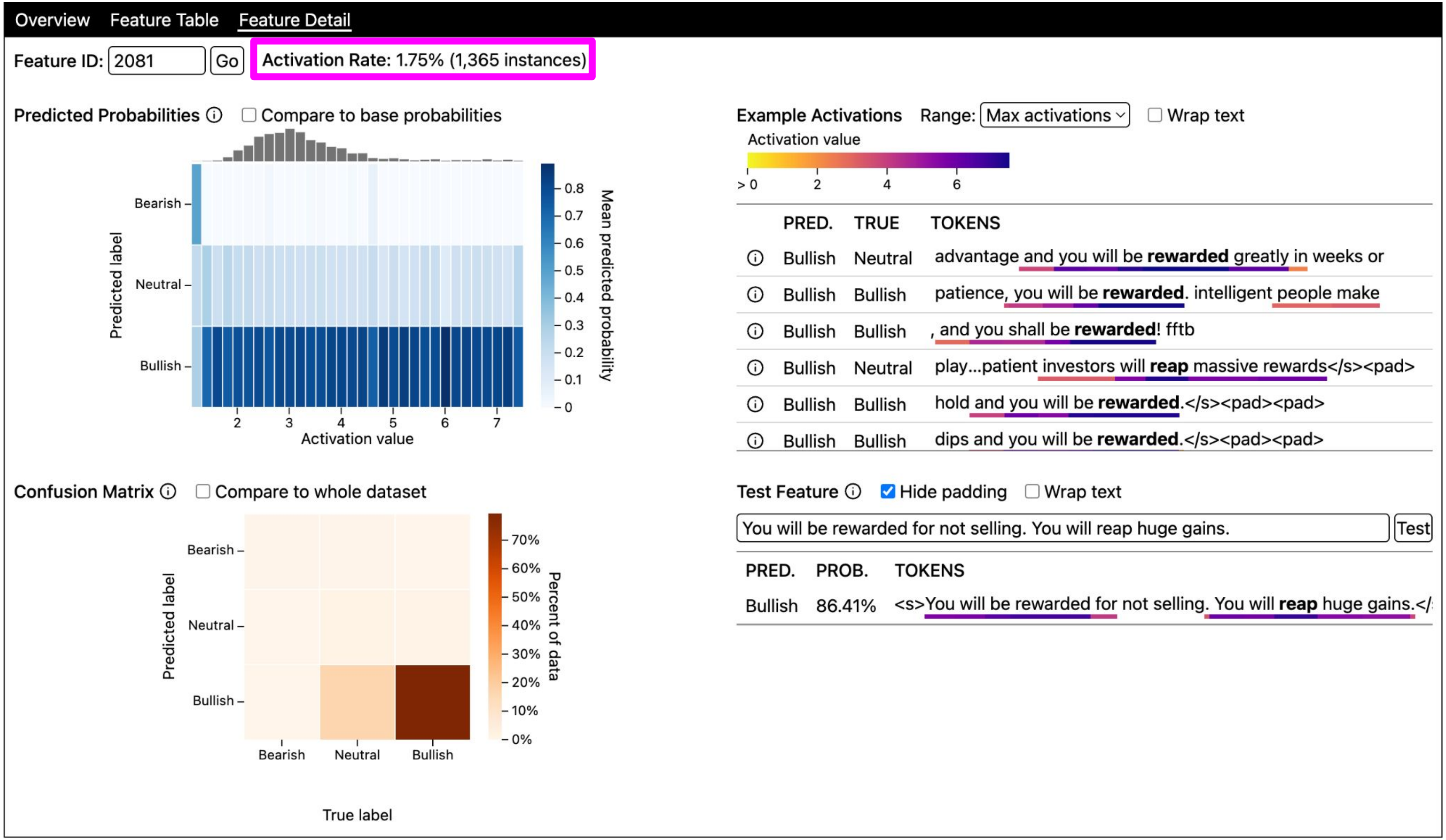
PRED.	TRUE	TOKENS
Bullish	Neutral	advantage and you will be **rewarded** greatly in weeks or
Bullish	Bullish	patience, you will be **rewarded**. intelligent people make
Bullish	Bullish	, and you shall be **rewarded**! fftb
Bullish	Neutral	play...patient investors will **reap** massive rewards</s><pad>
Bullish	Bullish	hold and you will be **rewarded**.</s><pad><pad>
Bullish	Bullish	dips and you will be **rewarded**.</s><pad><pad>

☑ Hide padding □ Wrap text

You will be rewarded for not selling. You will reap huge gains.

PRED.	PROB.	TOKENS
Bullish	86.41%	<s>You will be rewarded for not selling. You will **reap** huge gains.</

At the top left, there is the ID of the currently displayed feature.



Feature ID: 2081

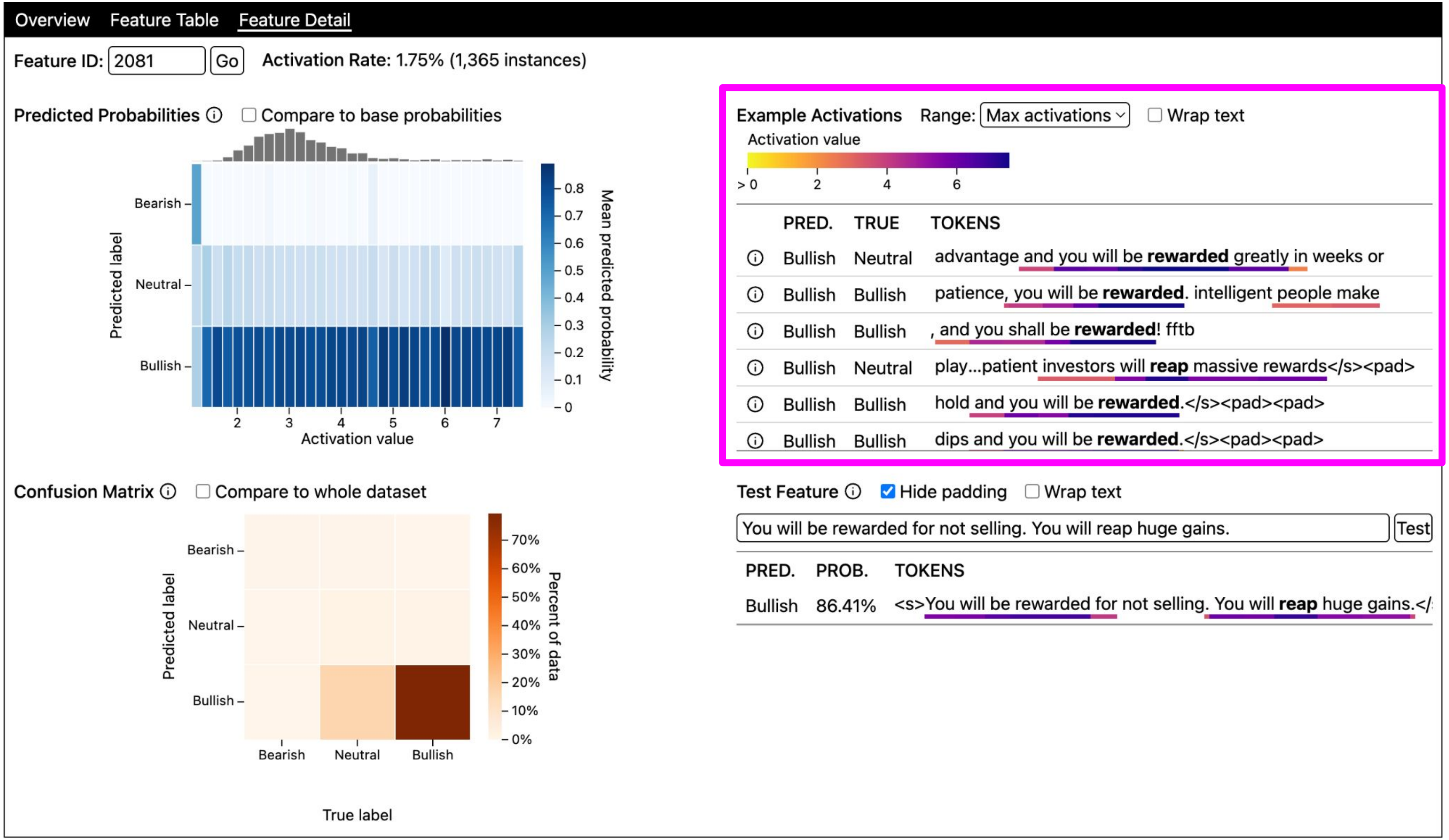
Activation Rate: 1.75% (1,365 instances)

	PRED.	TRUE	TOKENS
ⓘ	Bullish	Neutral	advantage and you will be **rewarded** greatly in weeks or
ⓘ	Bullish	Bullish	patience, you will be **rewarded**. intelligent people make
ⓘ	Bullish	Bullish	, and you shall be **rewarded**! fftb
ⓘ	Bullish	Neutral	play...patient investors will **reap** massive rewards</s><pad>
ⓘ	Bullish	Bullish	hold and you will be **rewarded**.</s><pad><pad>
ⓘ	Bullish	Bullish	dips and you will be **rewarded**.</s><pad><pad>

You will be rewarded for not selling. You will reap huge gains.

PRED.	PROB.	TOKENS
Bullish	86.41%	<s>You will be rewarded for not selling. You will **reap** huge gains.</

Next, there is the activation rate of this feature.



Feature ID: 2081

Activation Rate: 1.75% (1,365 instances)

□ Compare to base probabilities

□ Compare to whole dataset

Example Activations Range: Max activations □ Wrap text

	PRED.	TRUE	TOKENS
ⓘ	Bullish	Neutral	advantage and you will be **rewarded** greatly in weeks or
ⓘ	Bullish	Bullish	patience, you will be **rewarded**. intelligent people make
ⓘ	Bullish	Bullish	, and you shall be **rewarded**! fftb
ⓘ	Bullish	Neutral	play...patient investors will **reap** massive rewards</s><pad>
ⓘ	Bullish	Bullish	hold and you will be **rewarded**.</s><pad><pad>
ⓘ	Bullish	Bullish	dips and you will be **rewarded**.</s><pad><pad>

Test Feature ⓘ ☑ Hide padding □ Wrap text

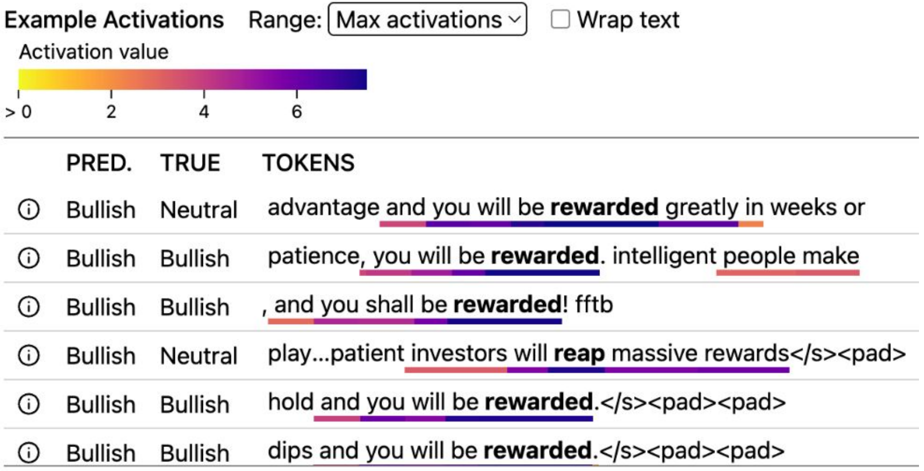
You will be rewarded for not selling. You will reap huge gains.

PRED.	PROB.	TOKENS
Bullish	86.41%	<s>You will be rewarded for not selling. You will **reap** huge gains.</

In the top right, we show snippets of the instances that activate the feature the most.

Example Activations

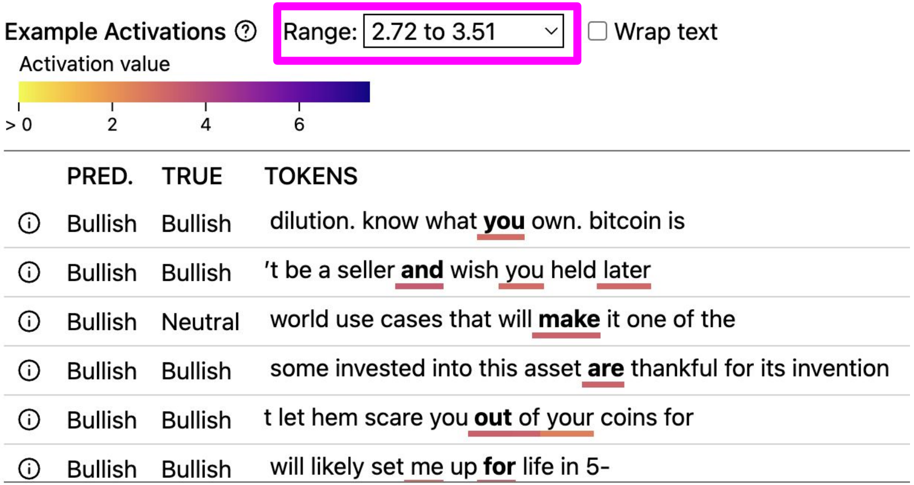
- The colored line beneath a token encodes the feature's activation value on that token.
- The token with the highest activation in the instance is also bolded.
- The 5 tokens before and after this token are provided as context.
- The table shows the predicted and true label for each instance.
- The information icon provides additional details about the instance.
- This table scrolls.



	PRED.	TRUE	TOKENS
ⓘ	Bullish	Neutral	advantage and you will be **rewarded** greatly in weeks or
ⓘ	Bullish	Bullish	patience, you will be **rewarded**. intelligent people make
ⓘ	Bullish	Bullish	, and you shall be **rewarded**! fftb
ⓘ	Bullish	Neutral	play...patient investors will **reap** massive rewards</s><pad>
ⓘ	Bullish	Bullish	hold and you will be **rewarded**.</s><pad><pad>
ⓘ	Bullish	Bullish	dips and you will be **rewarded**.</s><pad><pad>

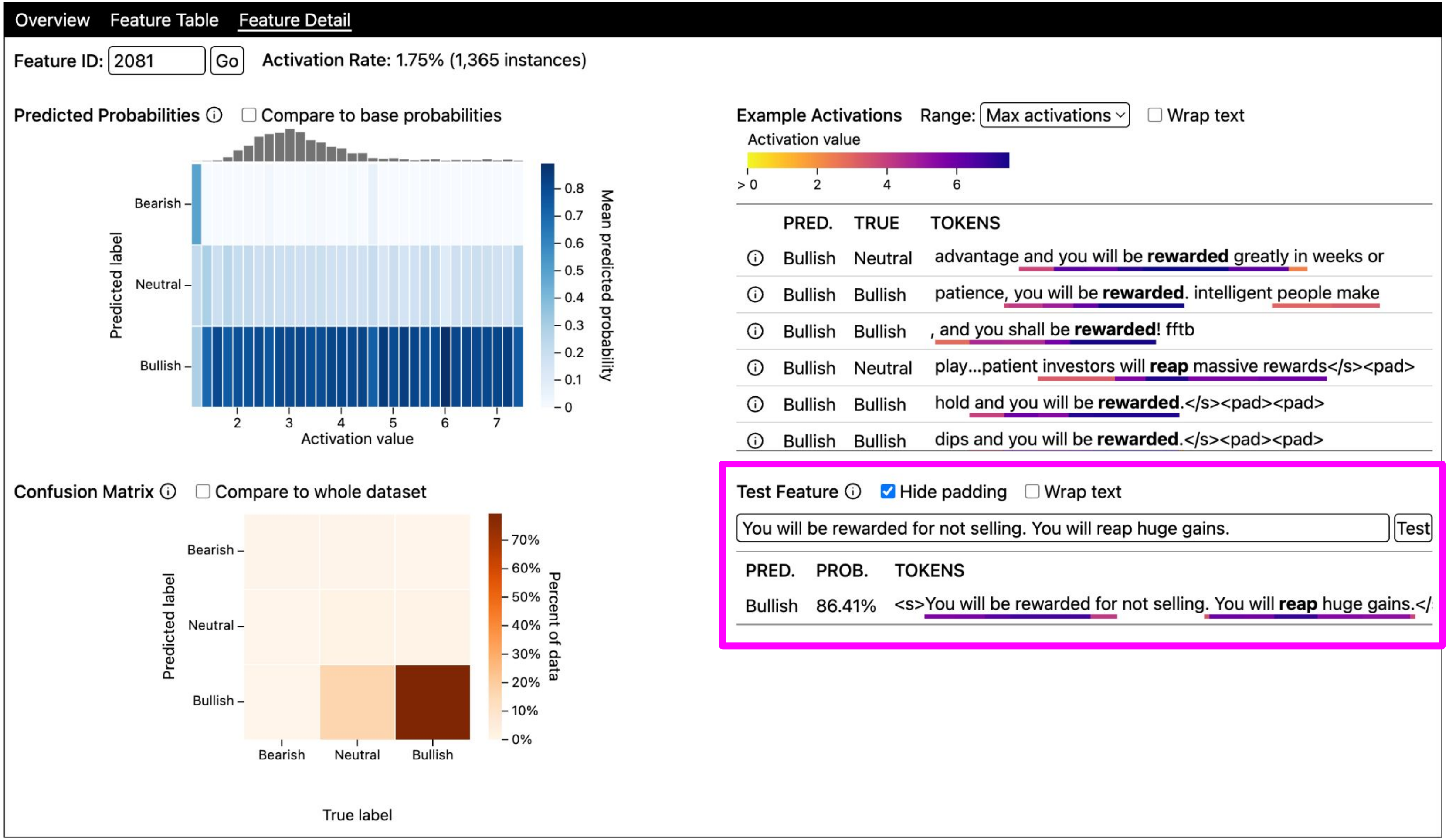
Example Activations

- By default, the instances with highest activations are shown.
- Adjusting the range lets you examples across the distribution of activation values.



Example Activations ⓘ Range: 2.72 to 3.51 ☐ Wrap text

	PRED.	TRUE	TOKENS
ⓘ	Bullish	Bullish	dilution. know what **you** own. bitcoin is
ⓘ	Bullish	Bullish	't be a seller **and** wish you held later
ⓘ	Bullish	Neutral	world use cases that will **make** it one of the
ⓘ	Bullish	Bullish	some invested into this asset **are** thankful for its invention
ⓘ	Bullish	Bullish	t let hem scare you **out** of your coins for
ⓘ	Bullish	Bullish	will likely set me up **for** life in 5-



Feature ID: 2081 Activation Rate: 1.75% (1,365 instances)

□ Compare to base probabilities

□ Compare to whole dataset

Example Activations Range: Max activations □ Wrap text

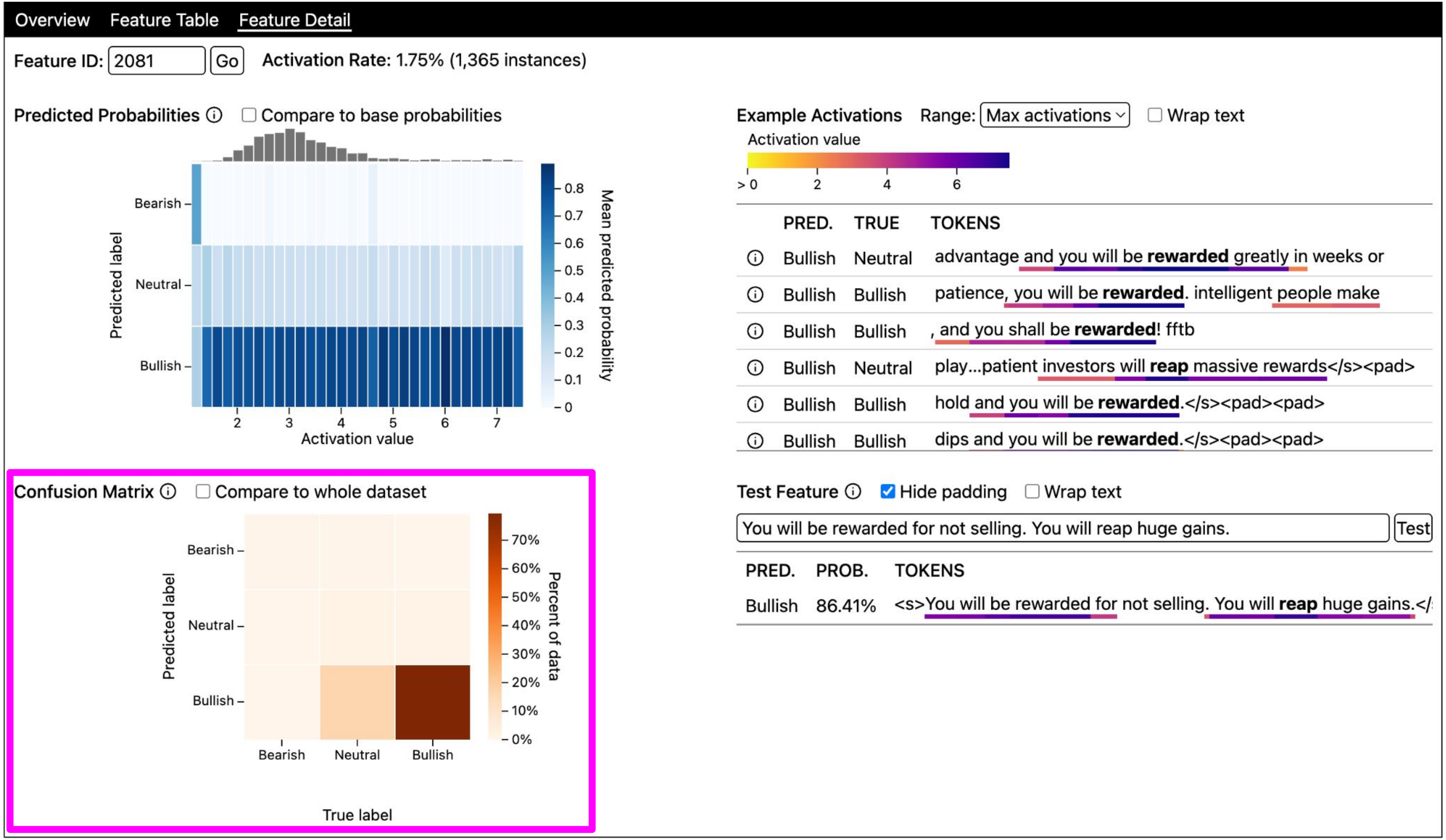
	PRED.	TRUE	TOKENS
ⓘ	Bullish	Neutral	advantage and you will be **rewarded** greatly in weeks or
ⓘ	Bullish	Bullish	patience, you will be **rewarded**. intelligent people make
ⓘ	Bullish	Bullish	, and you shall be **rewarded**! fftb
ⓘ	Bullish	Neutral	play...patient investors will **reap** massive rewards</s><pad>
ⓘ	Bullish	Bullish	hold and you will be **rewarded**.</s><pad><pad>
ⓘ	Bullish	Bullish	dips and you will be **rewarded**.</s><pad><pad>

Test Feature ⓘ ☑ Hide padding □ Wrap text

You will be rewarded for not selling. You will reap huge gains.

PRED.	PROB.	TOKENS
Bullish	86.41%	<s>You will be rewarded for not selling. You will **reap** huge gains.</

In the bottom right, you can enter your own text and check if it activates the feature.



Feature ID: 2081 Go Activation Rate: 1.75% (1,365 instances)

Predicted Probabilities ⓘ □ Compare to base probabilities

Example Activations Range: Max activations ▾ □ Wrap text

	PRED.	TRUE	TOKENS
ⓘ	Bullish	Neutral	advantage and you will be **rewarded** greatly in weeks or
ⓘ	Bullish	Bullish	patience, you will be **rewarded**. intelligent people make
ⓘ	Bullish	Bullish	, and you shall be **rewarded**! fftb
ⓘ	Bullish	Neutral	play...patient investors will **reap** massive rewards</s><pad>
ⓘ	Bullish	Bullish	hold and you will be **rewarded**.</s><pad><pad>
ⓘ	Bullish	Bullish	dips and you will be **rewarded**.</s><pad><pad>

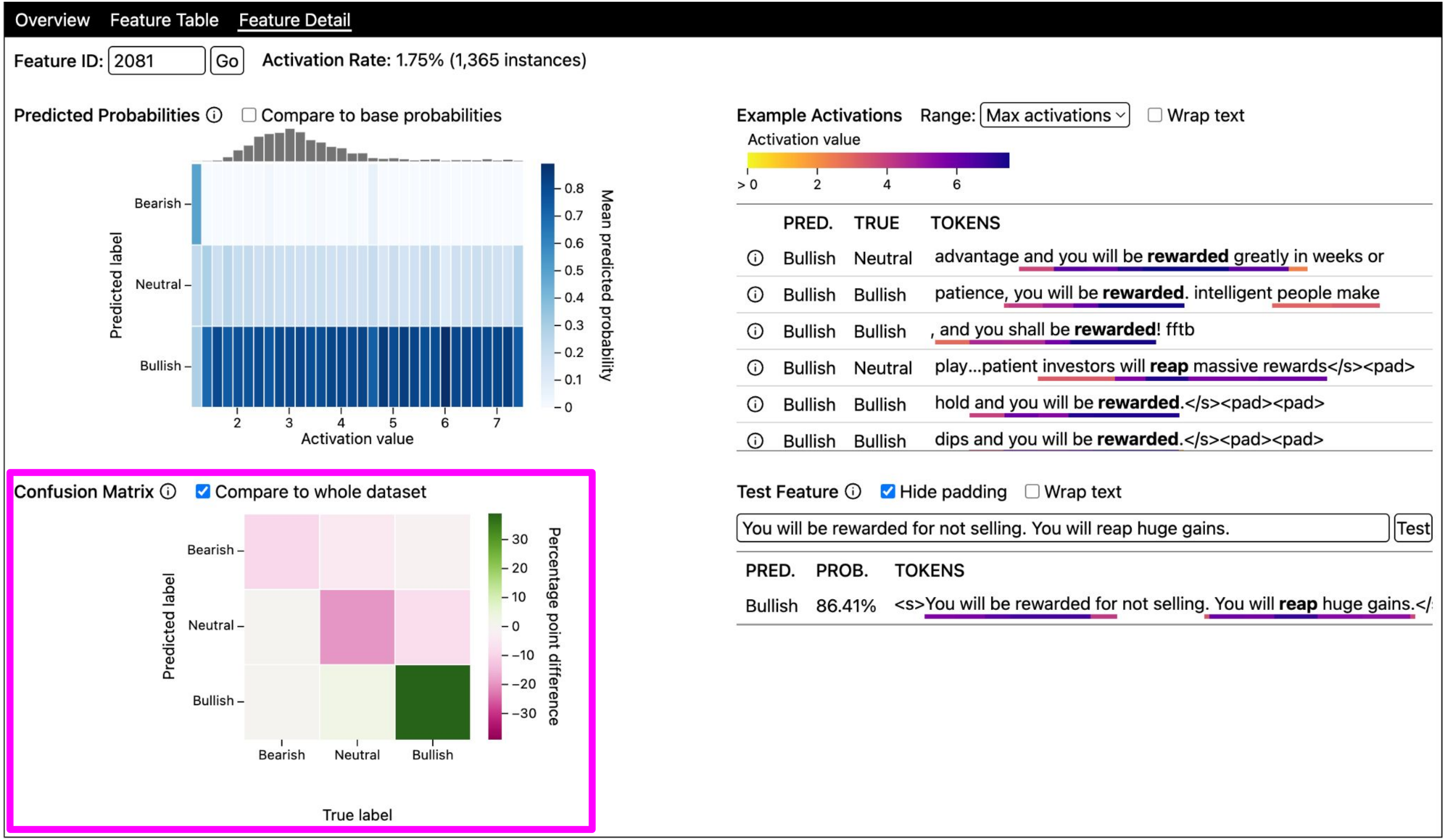
Test Feature ⓘ ☑ Hide padding □ Wrap text

You will be rewarded for not selling. You will reap huge gains. Test

PRED.	PROB.	TOKENS
Bullish	86.41%	<s>You will be rewarded for not selling. You will **reap** huge gains.</

Confusion Matrix ⓘ □ Compare to whole dataset

The confusion matrix shows the model’s performance on instances that activate this feature.



Feature ID: 2081 Activation Rate: 1.75% (1,365 instances)

Predicted Probabilities ⓘ □ Compare to base probabilities

Example Activations Range: Max activations □ Wrap text

PRED.	TRUE	TOKENS
Bullish	Neutral	advantage and you will be **rewarded** greatly in weeks or
Bullish	Bullish	patience, you will be **rewarded**. intelligent people make
Bullish	Bullish	, and you shall be **rewarded**! fftb
Bullish	Neutral	play...patient investors will **reap** massive rewards</s><pad>
Bullish	Bullish	hold and you will be **rewarded**.</s><pad><pad>
Bullish	Bullish	dips and you will be **rewarded**.</s><pad><pad>

Confusion Matrix ⓘ ☑ Compare to whole dataset

Test Feature ⓘ ☑ Hide padding □ Wrap text

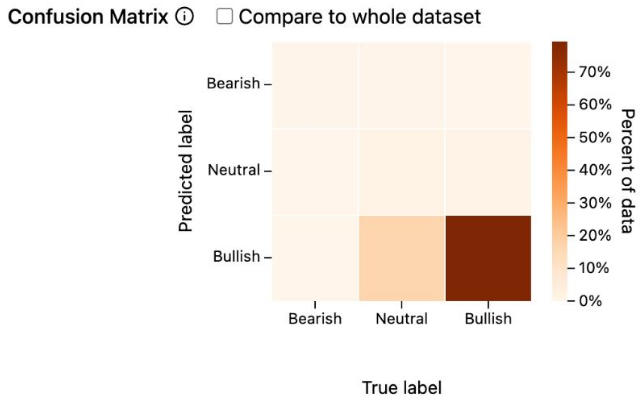
You will be rewarded for not selling. You will reap huge gains.

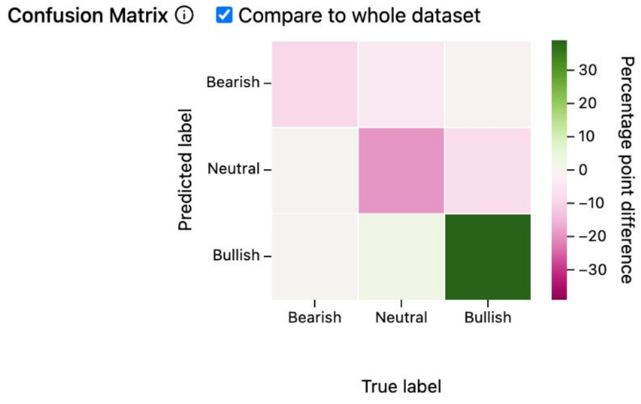
PRED.	PROB.	TOKENS
Bullish	86.41%	<s>You will be rewarded for not selling. You will **reap** huge gains.</

Checking “Compare to whole dataset” compares this confusion matrix to the one for all instances.

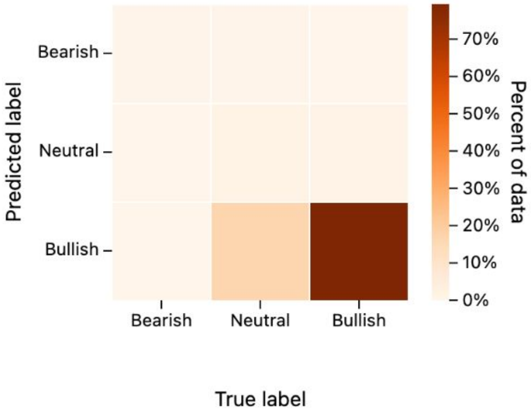
Confusion matrix comparison

- How do the model's predictions on the instances that activate this feature compare to its predictions on the whole dataset?
- Calculate the percentage point difference between the cells of this confusion matrix and the one for all instances.
- Ex: The model predicts Bullish more often on the instances that activate this feature than on the whole dataset.





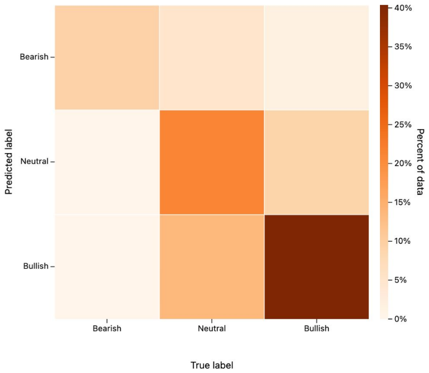
Confusion matrix from the Feature Detail tab.



Calculated from instances that activate the feature.

−

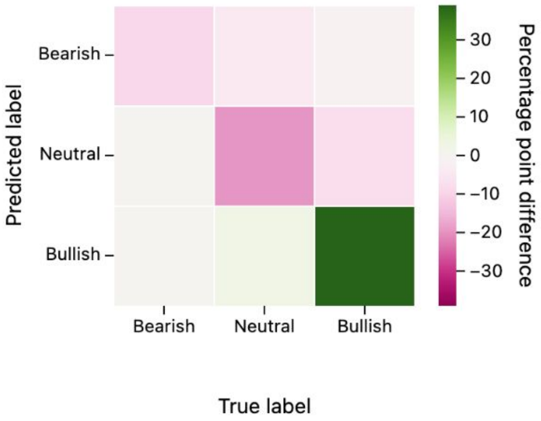
Confusion matrix from the Overview tab.

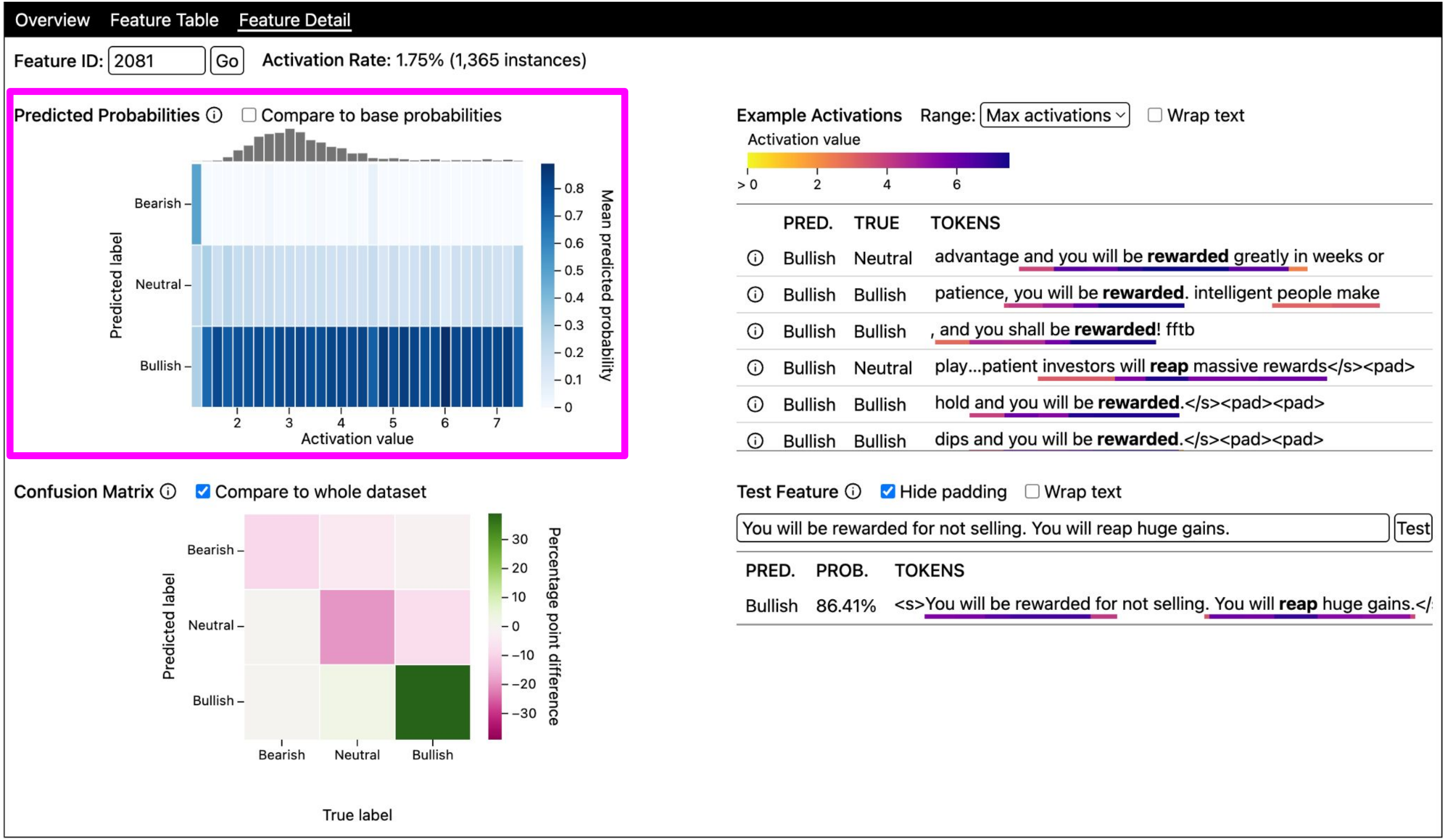


Calculated from all instances

=

Comparison to whole dataset





Feature ID: 2081 Go Activation Rate: 1.75% (1,365 instances)

Predicted Probabilities ⓘ ☐ Compare to base probabilities

Example Activations Range: Max activations ☐ Wrap text

	PRED.	TRUE	TOKENS
ⓘ	Bullish	Neutral	advantage and you will be **rewarded** greatly in weeks or
ⓘ	Bullish	Bullish	patience, you will be **rewarded**. intelligent people make
ⓘ	Bullish	Bullish	, and you shall be **rewarded**! fftb
ⓘ	Bullish	Neutral	play...patient investors will **reap** massive rewards</s><pad>
ⓘ	Bullish	Bullish	hold and you will be **rewarded**.</s><pad><pad>
ⓘ	Bullish	Bullish	dips and you will be **rewarded**.</s><pad><pad>

Confusion Matrix ⓘ ☑ Compare to whole dataset

Test Feature ⓘ ☑ Hide padding ☐ Wrap text

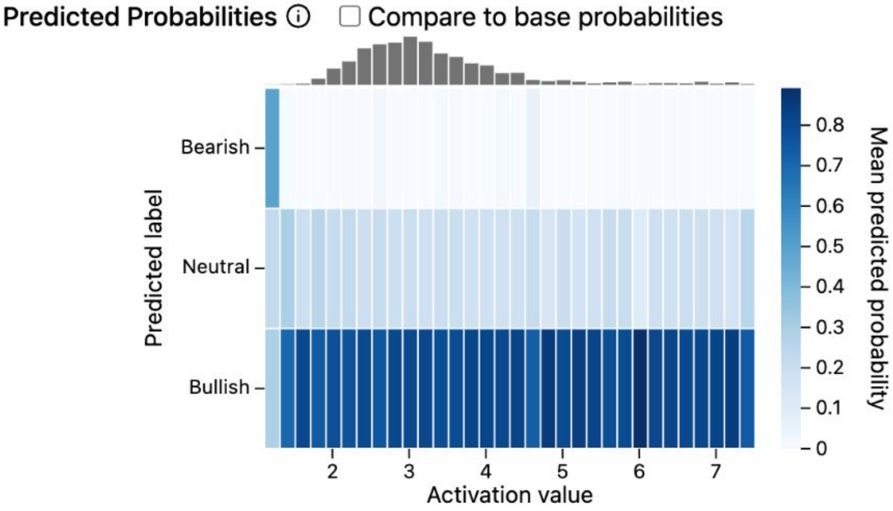
You will be rewarded for not selling. You will reap huge gains. Test

PRED.	PROB.	TOKENS
Bullish	86.41%	<s>You will be rewarded for not selling. You will **reap** huge gains.</

The top left shows the model's predicted probabilities on the instances that activate this feature.

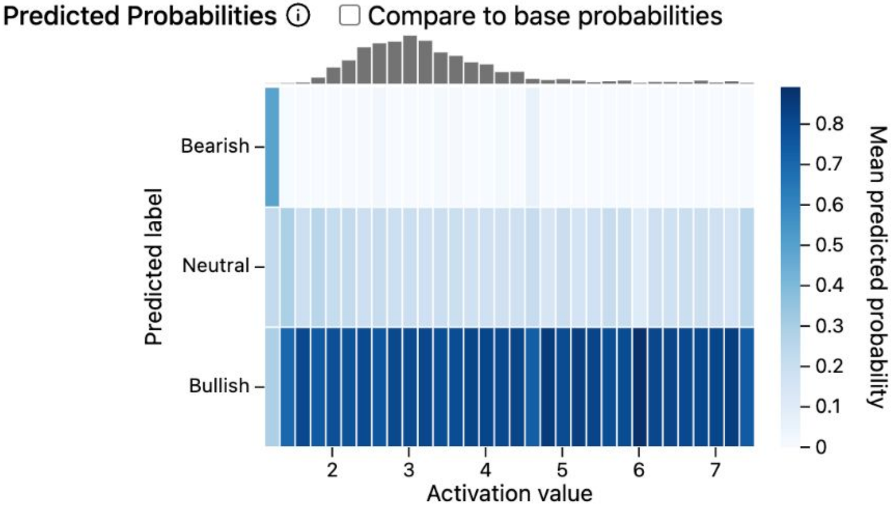
Predicted Probabilities

- What is the relationship between the feature's activation value and the model's predictions?
- Bin the instances that activate the feature by their maximum activation value
  - Multiple tokens in an instance can cause the feature to activate.
  - Across all of the tokens in the instance, take the maximum activation value.
- Calculate the model's mean predicted class probabilities for each bin.



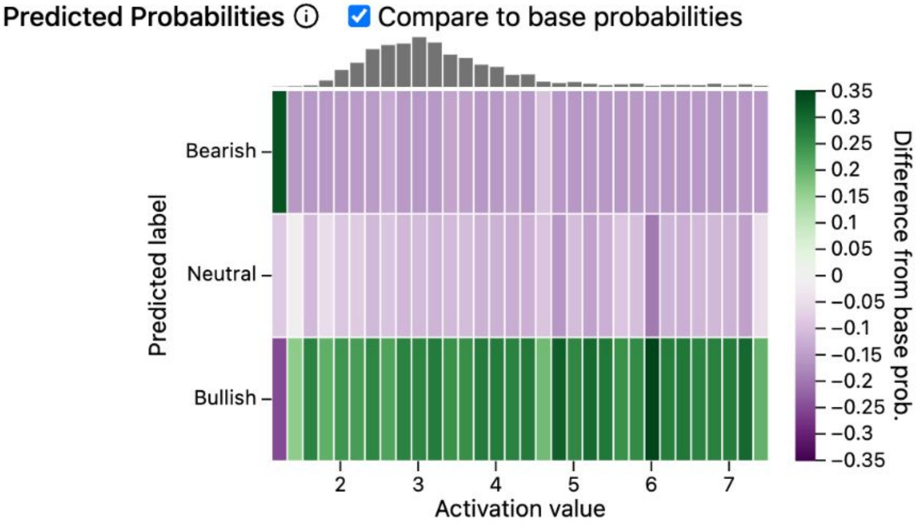
Predicted Probabilities

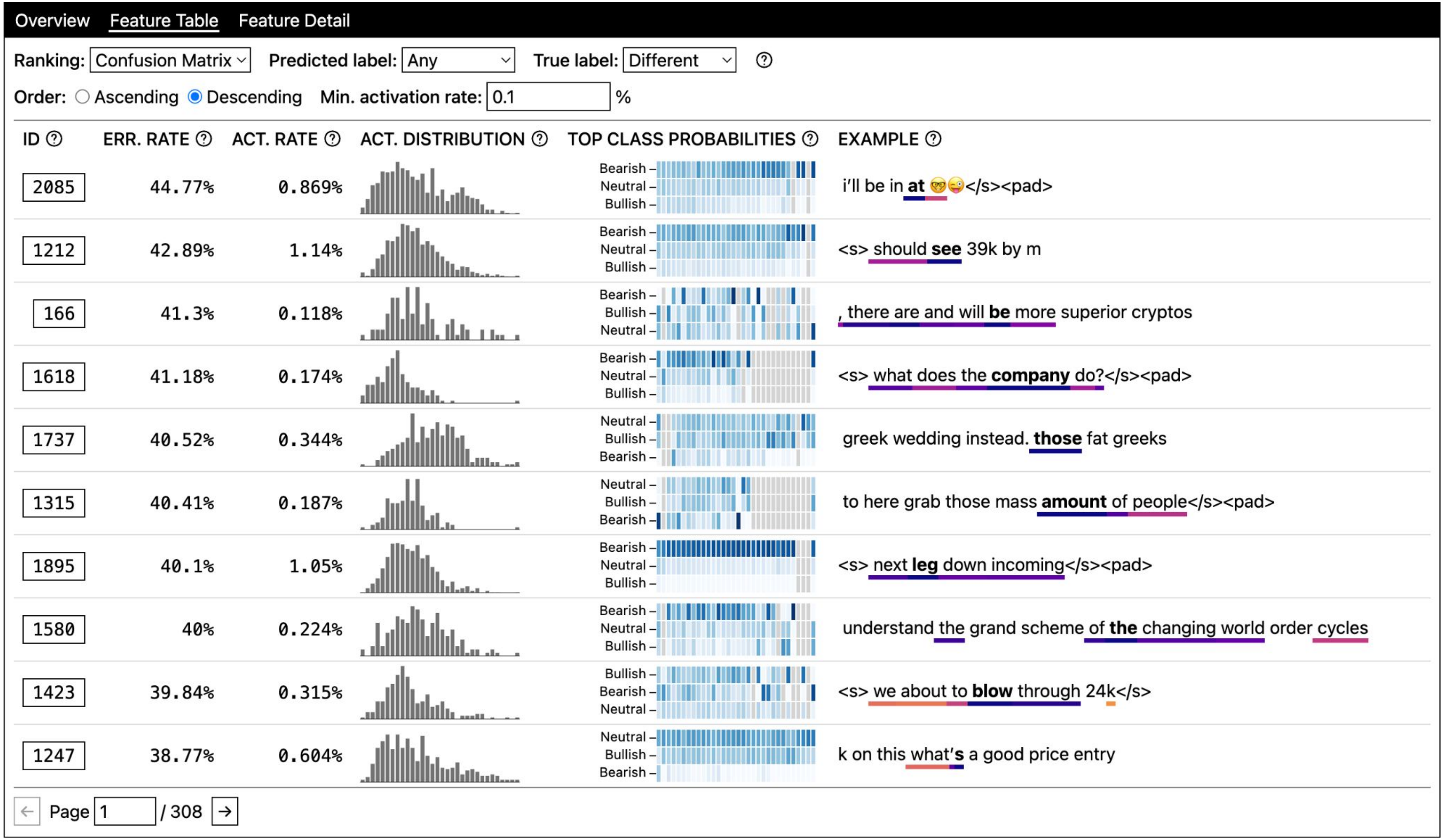
- Visualized with a heatmap.
- Each row represents a class.
- Each column represents a range of activation values.
- The color of a cell encodes model's mean predicted probability for the given class for instances that activate the feature in the given range.
- This histogram above the heatmap shows the number of instances in each activation value bin.
- Ex: The model assigns a high probability to Bullish for all activation value bins except the lowest.



Predicted Probabilities

- Checking “Compare to base probabilities” subtracts each row by the mean predicted probability for that class across all instances.
- This highlights how the model’s predicted probabilities on the instances that activate this feature differs from its predicted probabilities on the entire dataset.
- Ex: the model assigns higher probability to Bullish for these instances than it does for the dataset as a whole.





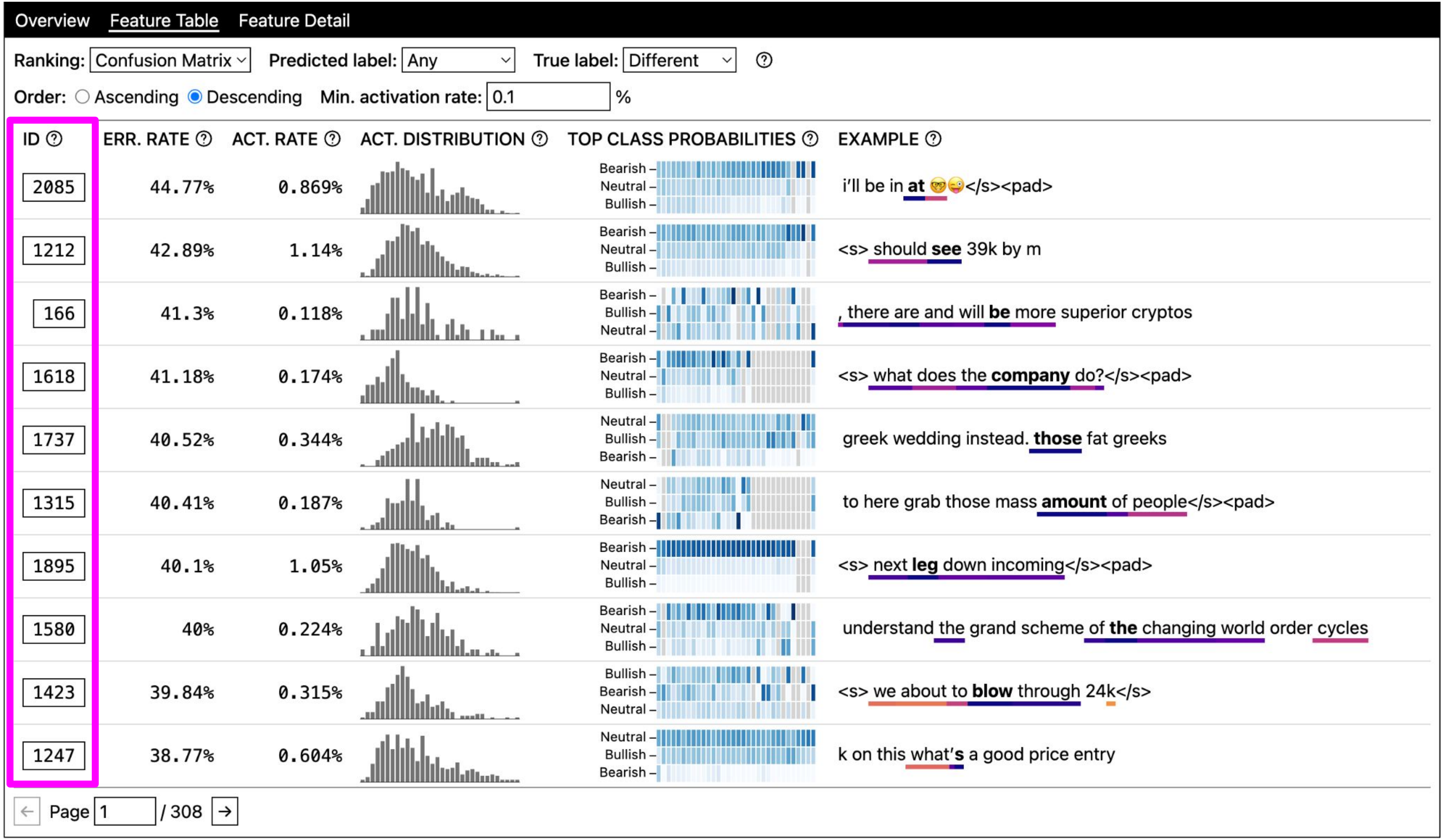
Ranking: Confusion Matrix ⌄ Predicted label: Any ⌄ True label: Different ⌄ ?

Order: ○ Ascending ● Descending Min. activation rate: 0.1 %

ID ?	ERR. RATE ?	ACT. RATE ?	ACT. DISTRIBUTION ?	TOP CLASS PROBABILITIES ?	EXAMPLE ?
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

Finally, we will cover the Feature Table tab.



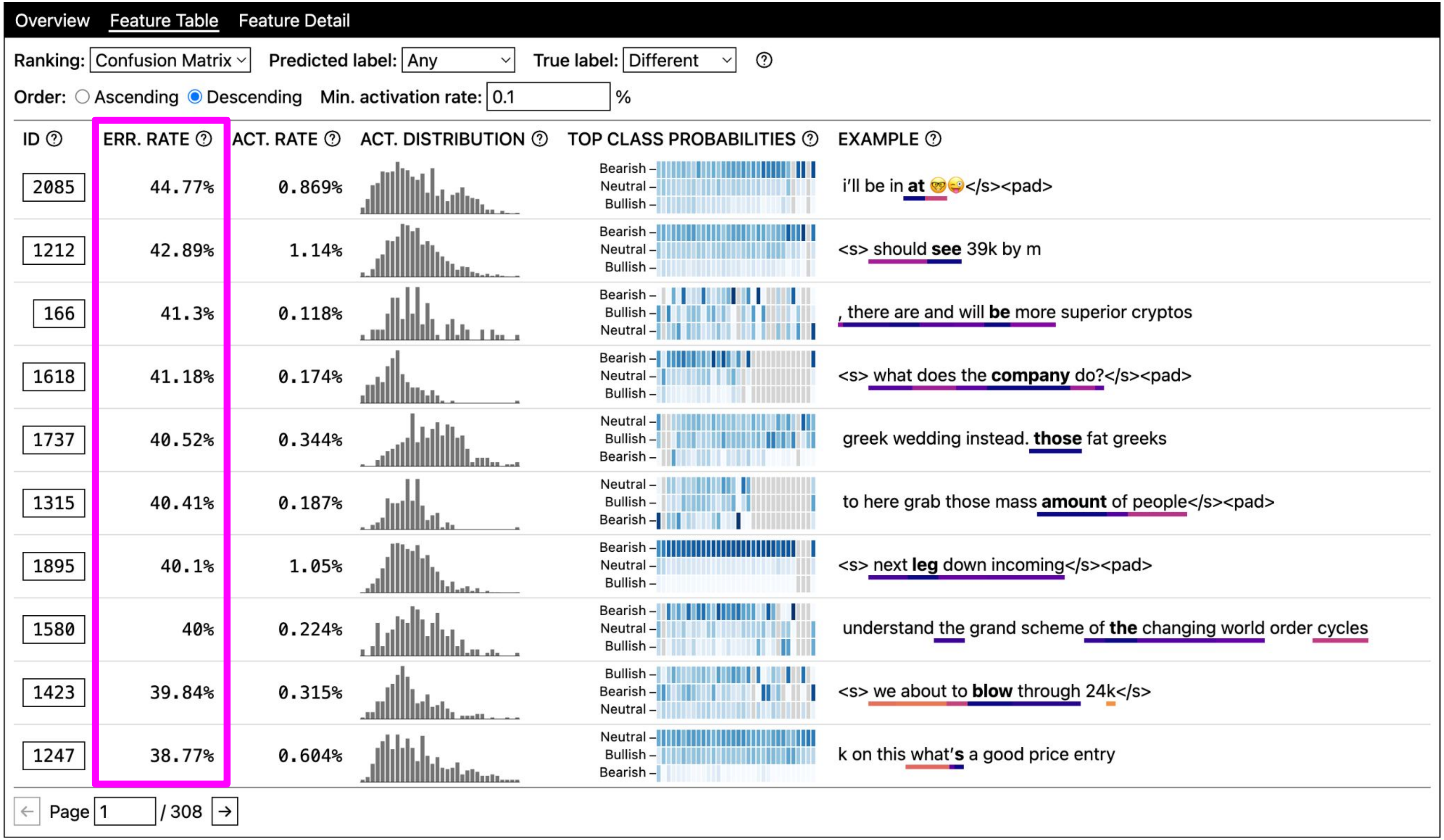
Ranking: Confusion Matrix Predicted label: Any True label: Different ⓘ

Order: ○ Ascending ● Descending Min. activation rate: 0.1 %

ID ⓘ	ERR. RATE ⓘ	ACT. RATE ⓘ	ACT. DISTRIBUTION ⓘ	TOP CLASS PROBABILITIES ⓘ	EXAMPLE ⓘ
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

The 1st column contains the ID of the feature. Clicking the ID goes to the Feature Detail tab.



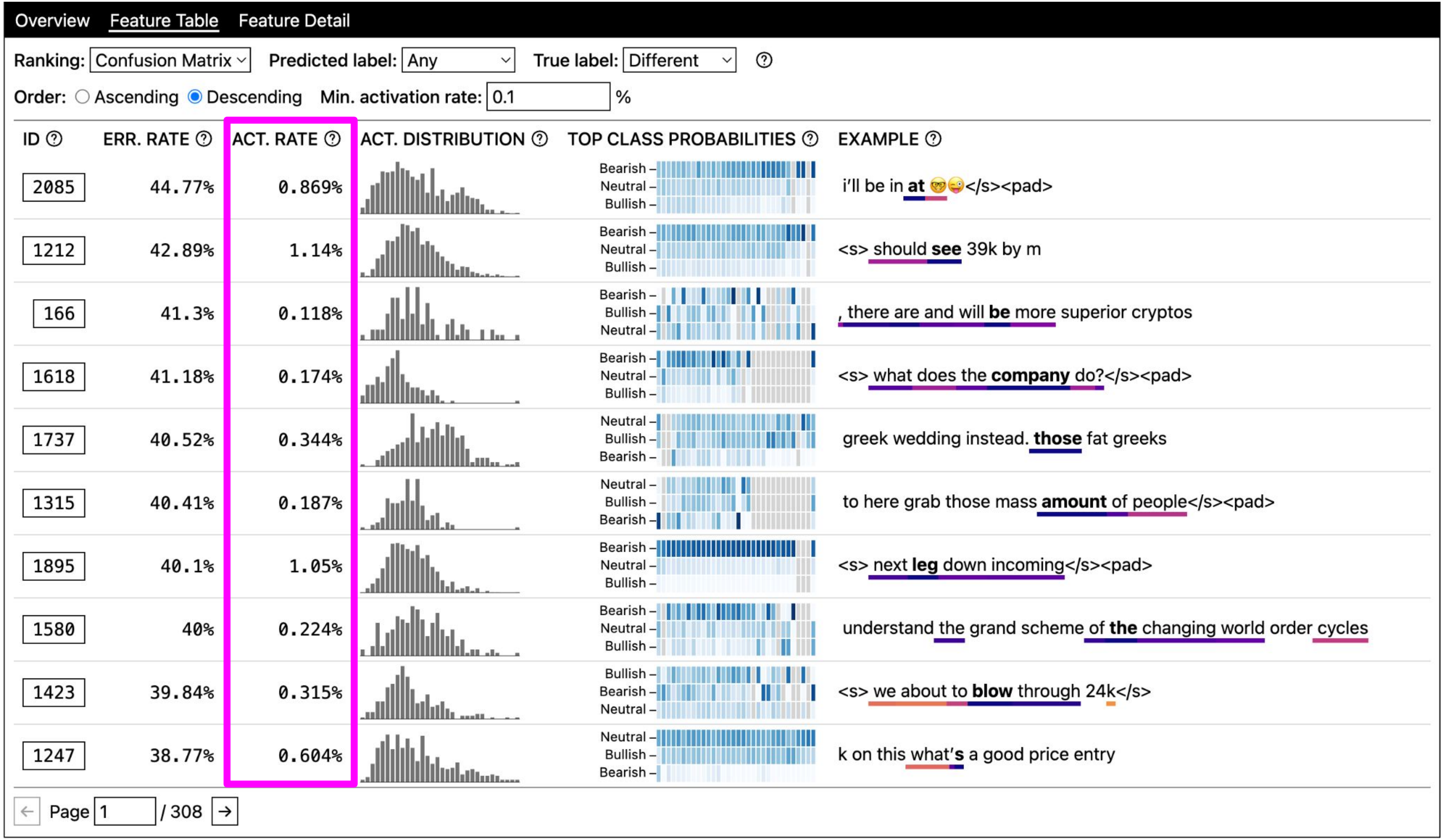
Ranking: Confusion Matrix ⌄ Predicted label: Any ⌄ True label: Different ⌄ ⓘ

Order: ○ Ascending ● Descending Min. activation rate: 0.1 %

ID ⓘ	ERR. RATE ⓘ	ACT. RATE ⓘ	ACT. DISTRIBUTION ⓘ	TOP CLASS PROBABILITIES ⓘ	EXAMPLE ⓘ
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

The 2nd column shows the model's error rate on the instances that activate the feature.



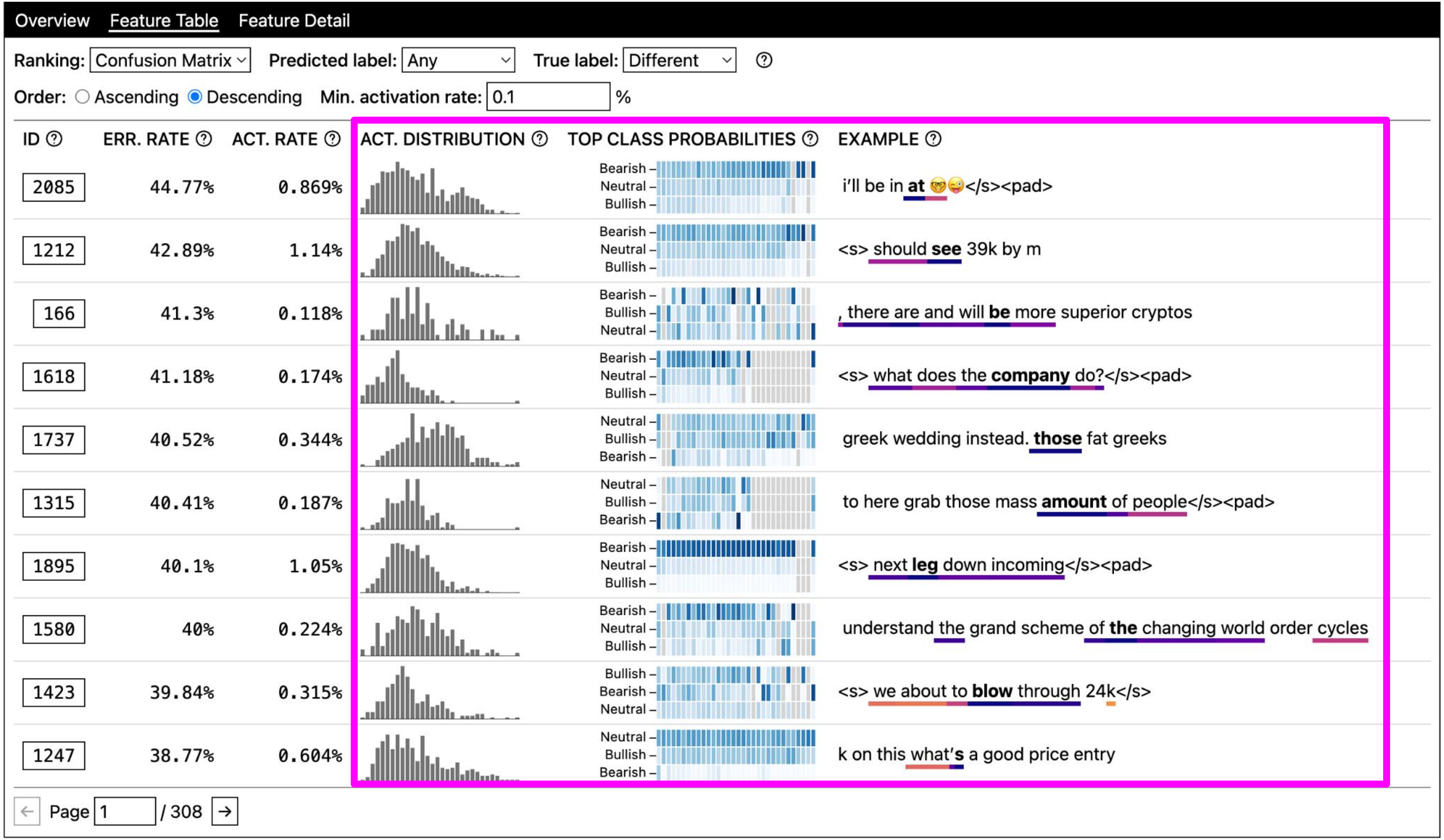
Ranking: Confusion Matrix ⌄ Predicted label: Any ⌄ True label: Different ⌄ ?

Order: ○ Ascending ● Descending Min. activation rate: 0.1 %

ID ?	ERR. RATE ?	ACT. RATE ?	ACT. DISTRIBUTION ?	TOP CLASS PROBABILITIES ?	EXAMPLE ?
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

The 3rd column shows the feature's activation rate.



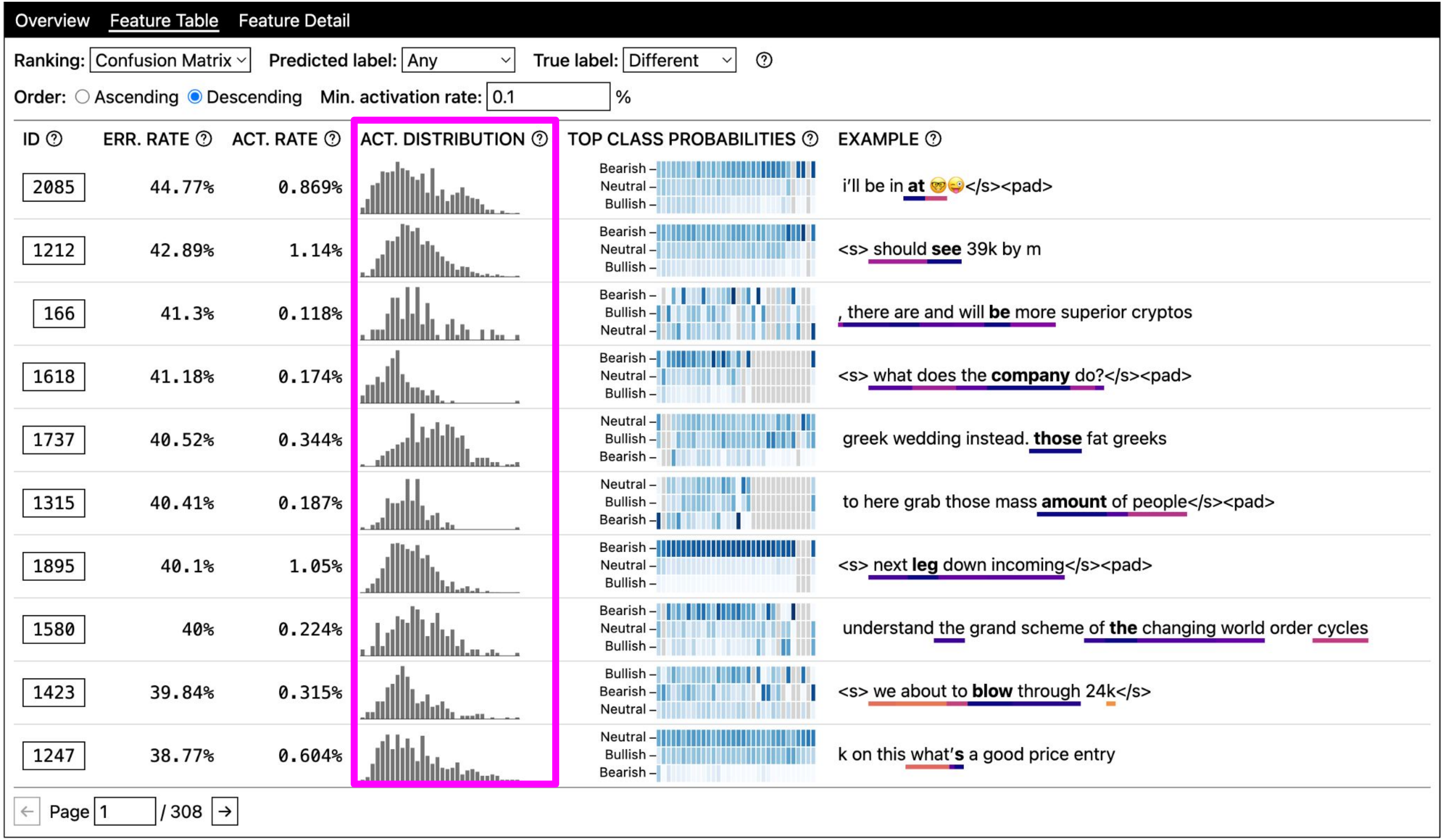
Ranking: Confusion Matrix ⌄ Predicted label: Any ⌄ True label: Different ⌄ ⓘ

Order: ○ Ascending ● Descending Min. activation rate: 0.1 %

ID ⓘ	ERR. RATE ⓘ	ACT. RATE ⓘ	ACT. DISTRIBUTION ⓘ	TOP CLASS PROBABILITIES ⓘ	EXAMPLE ⓘ
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

The remaining columns are visualizations that are also on the Feature Detail tab.



Overview | Feature Table | Feature Detail

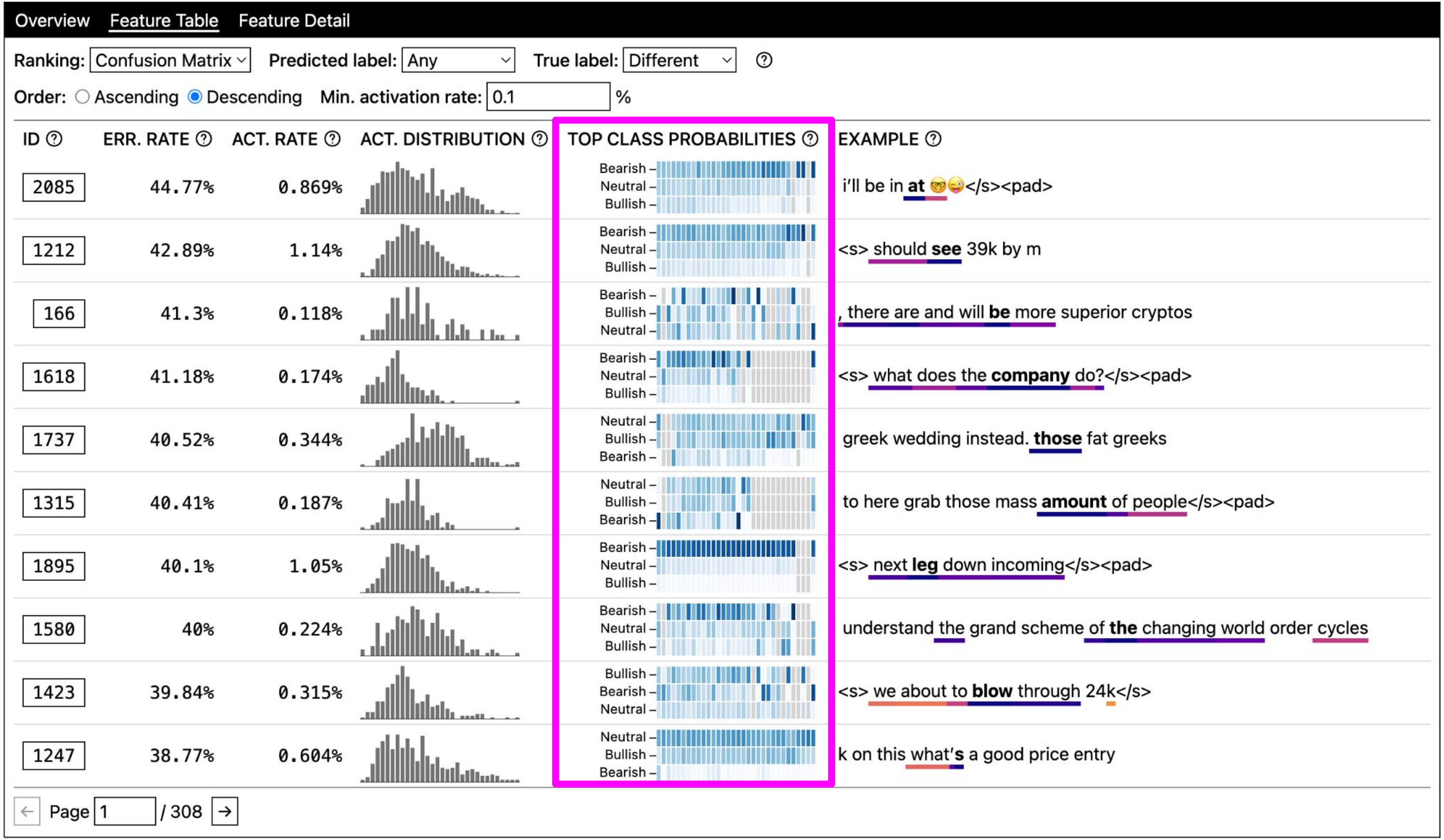
Ranking: Confusion Matrix   Predicted label: Any   True label: Different   ⓘ

Order: ○ Ascending ● Descending   Min. activation rate: 0.1 %

ID ⓘ	ERR. RATE ⓘ	ACT. RATE ⓘ	ACT. DISTRIBUTION ⓘ	TOP CLASS PROBABILITIES ⓘ	EXAMPLE ⓘ
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

The 4th column shows a histogram of the activation values.



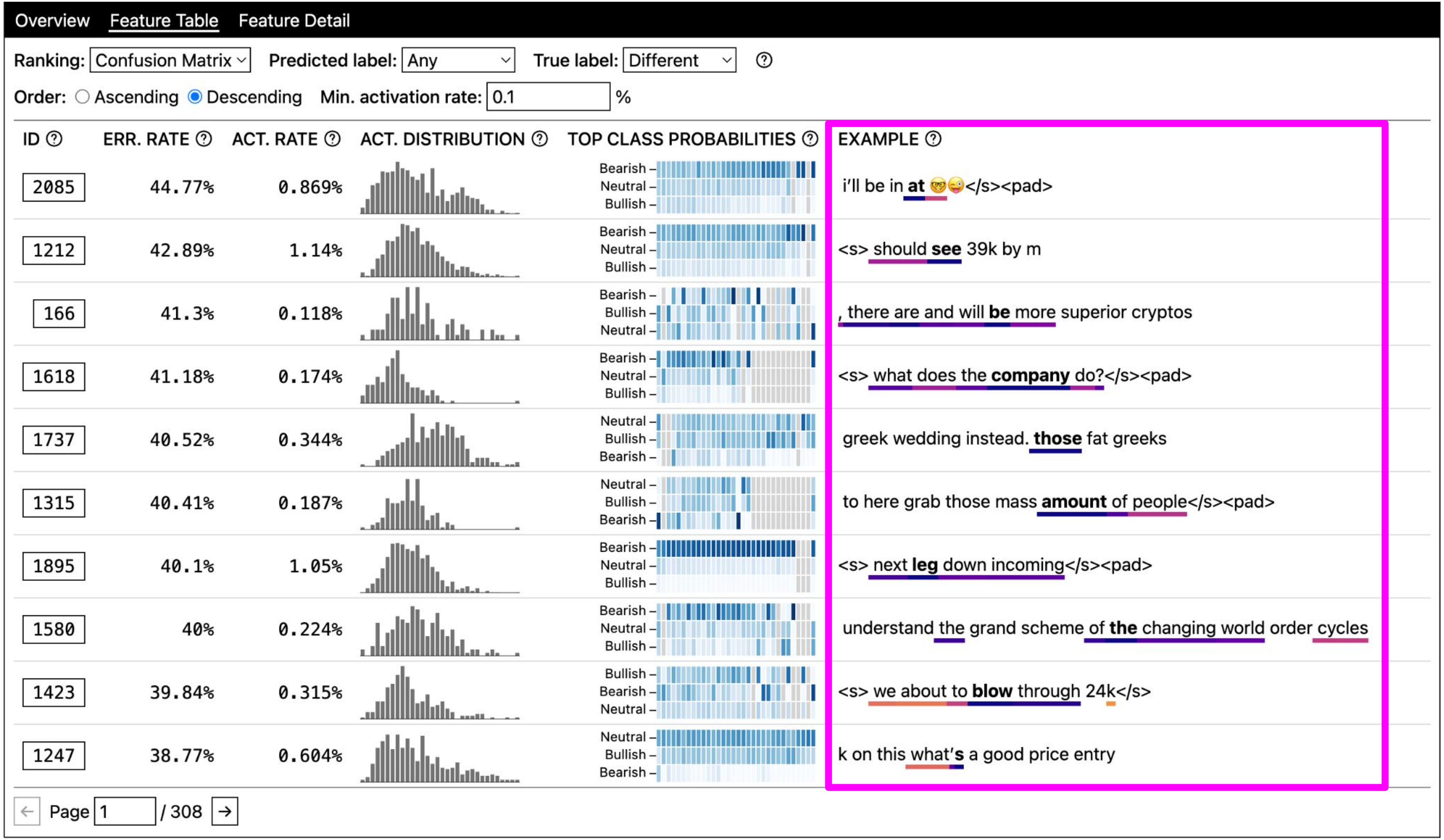
Ranking: Confusion Matrix Predicted label: Any True label: Different

Order: ○ Ascending ● Descending Min. activation rate: 0.1 %

ID	ERR. RATE	ACT. RATE	ACT. DISTRIBUTION	TOP CLASS PROBABILITIES	EXAMPLE
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

The 5th column shows the model's predicted probabilities for the top 3 classes.



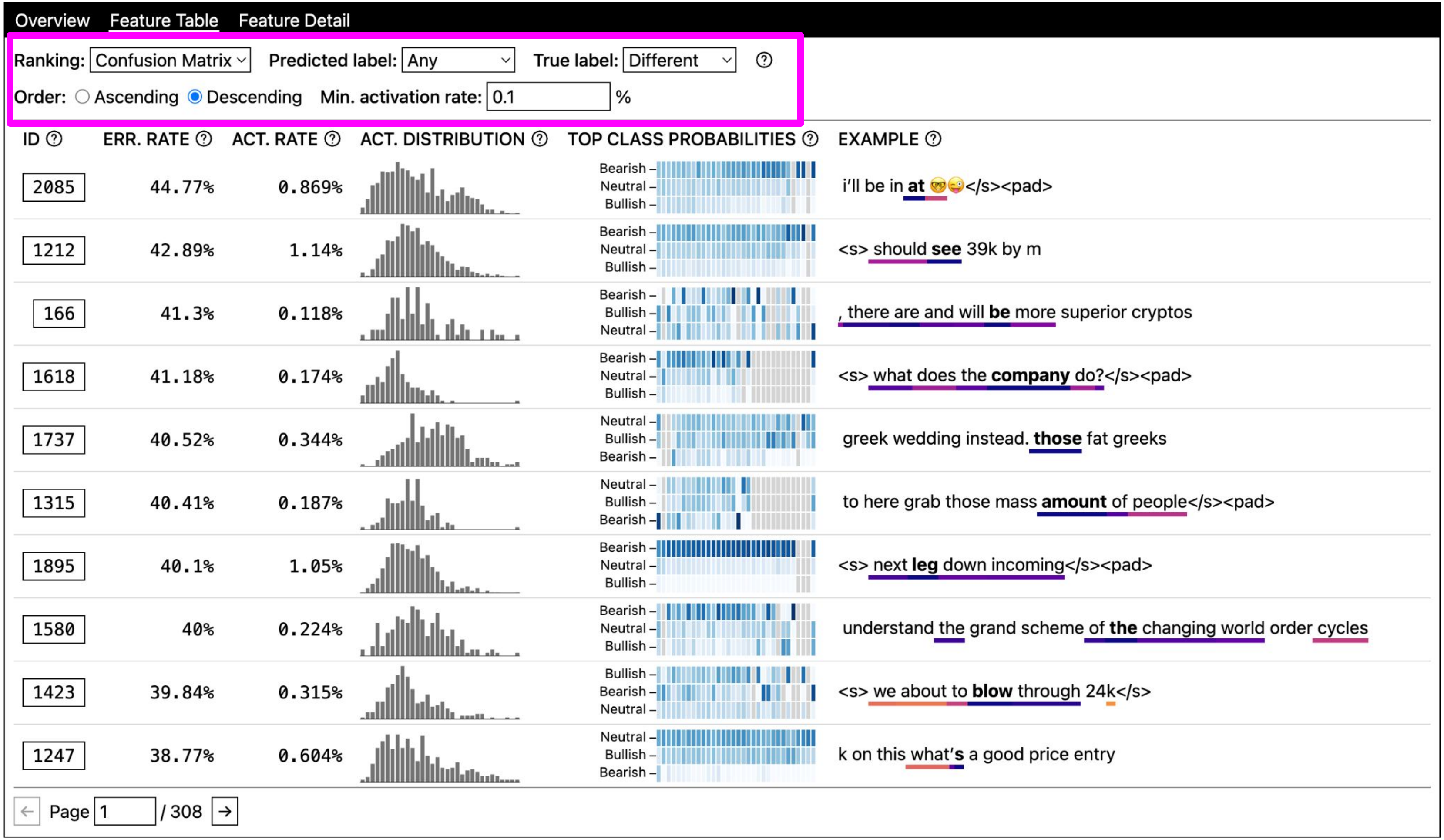
Ranking: Confusion Matrix   Predicted label: Any   True label: Different

Order: ○ Ascending ● Descending   Min. activation rate: 0.1 %

ID	ERR. RATE	ACT. RATE	ACT. DISTRIBUTION	TOP CLASS PROBABILITIES	EXAMPLE
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

Page 1 / 308

The last column shows the sequence of tokens that maximally activates the feature.



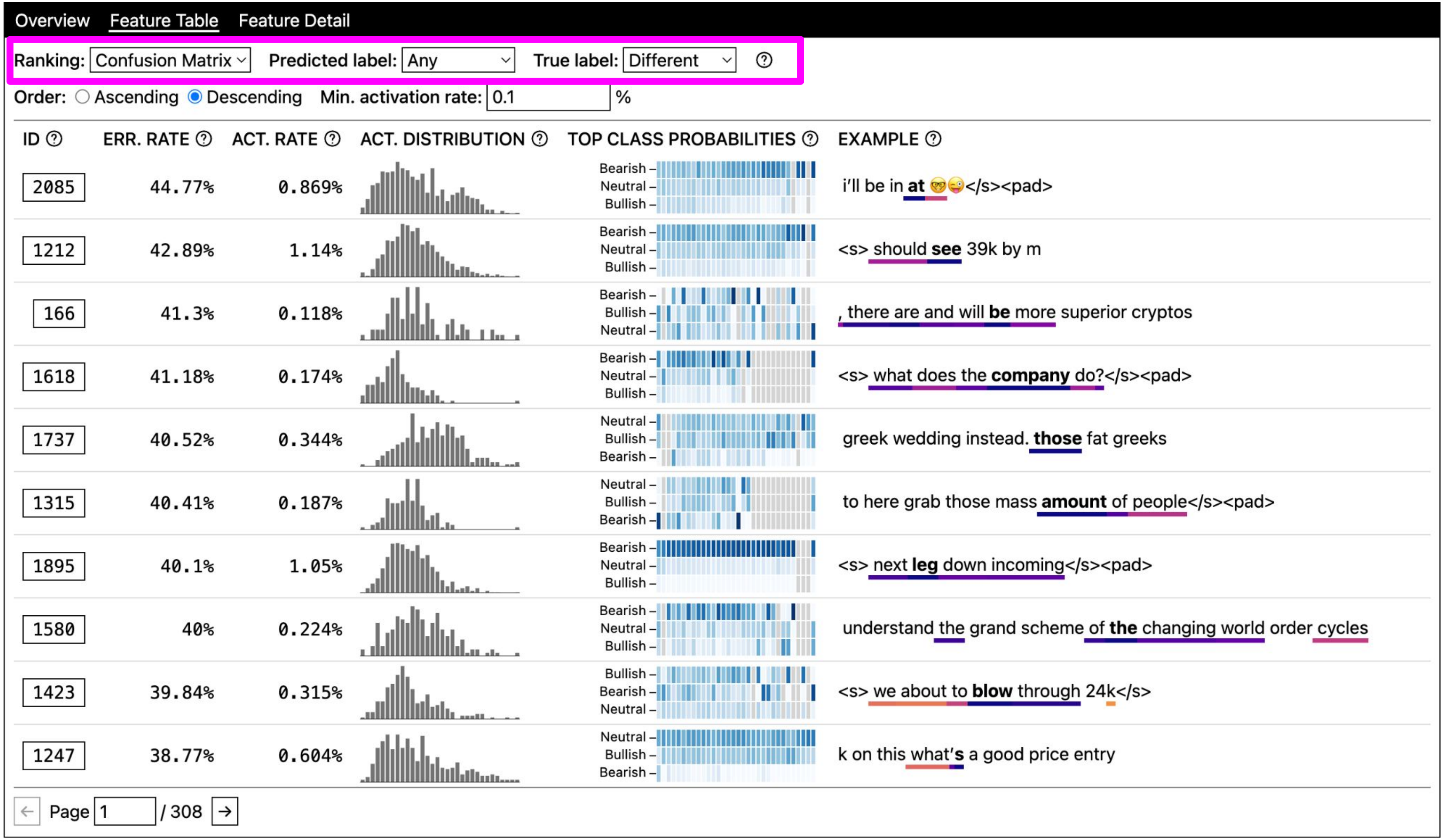
Ranking: Confusion Matrix ⌄ Predicted label: Any ⌄ True label: Different ⌄ ⓘ

Order: ○ Ascending ● Descending Min. activation rate: 0.1 %

ID ⓘ	ERR. RATE ⓘ	ACT. RATE ⓘ	ACT. DISTRIBUTION ⓘ	TOP CLASS PROBABILITIES ⓘ	EXAMPLE ⓘ
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

At the top of the tab, you can control how the features are sorted and filtered.



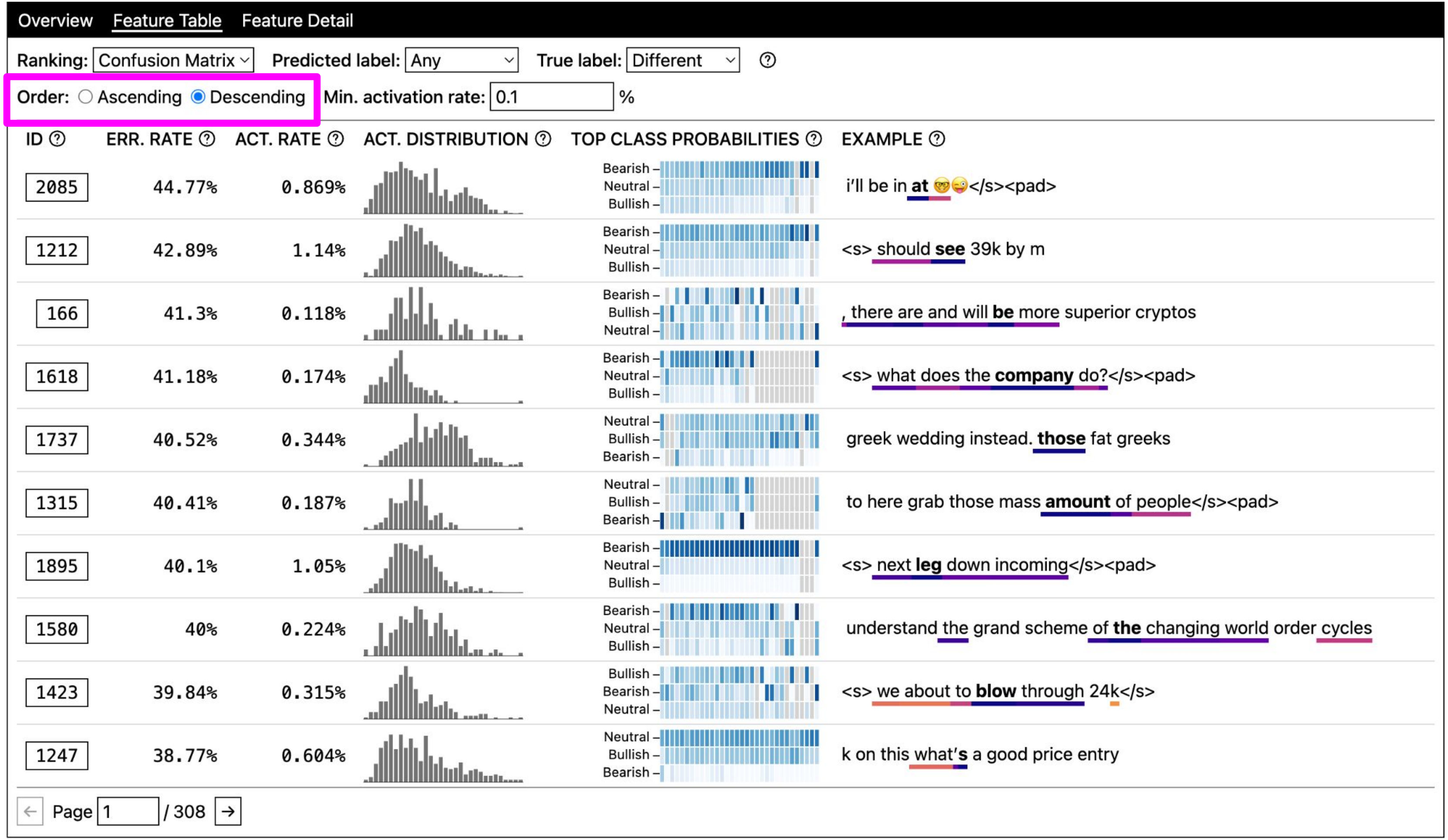
Ranking: Confusion Matrix ⌄ Predicted label: Any ⌄ True label: Different ⌄ ⓘ

Order: ○ Ascending ● Descending Min. activation rate: 0.1 %

ID ⓘ	ERR. RATE ⓘ	ACT. RATE ⓘ	ACT. DISTRIBUTION ⓘ	TOP CLASS PROBABILITIES ⓘ	EXAMPLE ⓘ
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

The first row of controls sets ranking method.



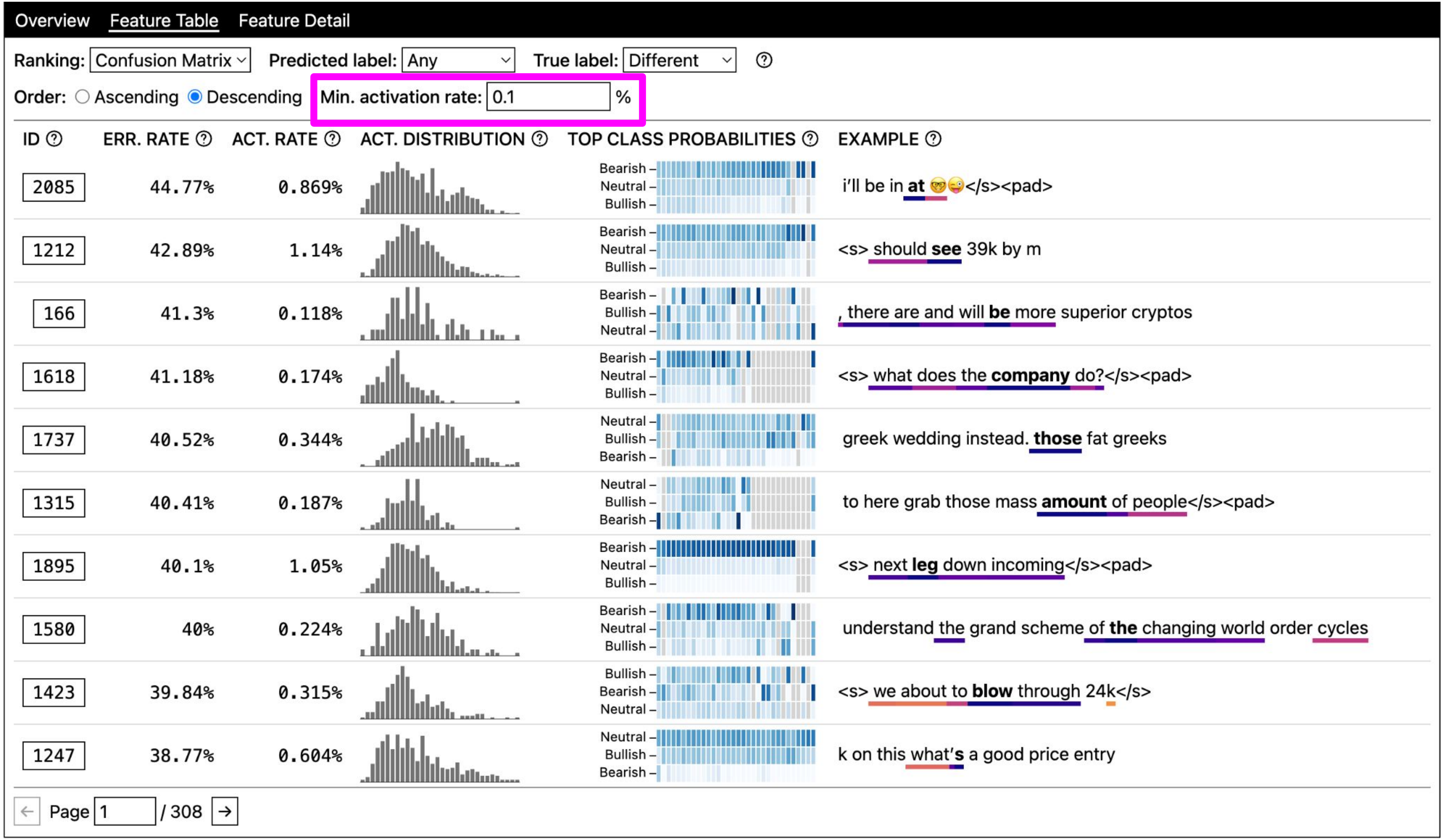
Ranking: Confusion Matrix Predicted label: Any True label: Different ⓘ

Order: ○ Ascending ● Descending Min. activation rate: 0.1 %

ID ⓘ	ERR. RATE ⓘ	ACT. RATE ⓘ	ACT. DISTRIBUTION ⓘ	TOP CLASS PROBABILITIES ⓘ	EXAMPLE ⓘ
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

In the second row, you can choose between ranking in ascending or descending order.



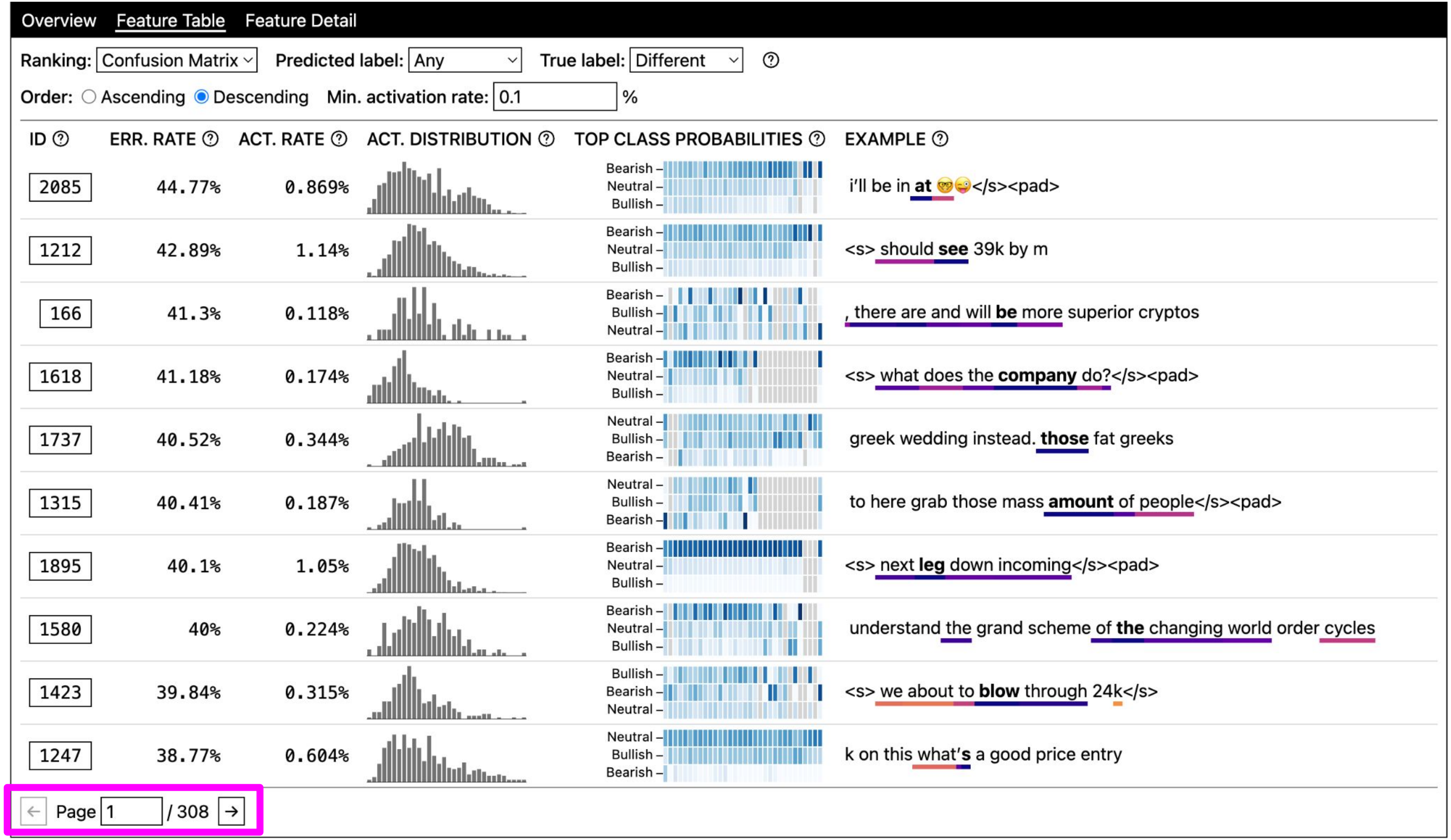
Ranking: Confusion Matrix Predicted label: Any True label: Different ⓘ

Order: ○ Ascending ● Descending Min. activation rate: 0.1 %

ID ⓘ	ERR. RATE ⓘ	ACT. RATE ⓘ	ACT. DISTRIBUTION ⓘ	TOP CLASS PROBABILITIES ⓘ	EXAMPLE ⓘ
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

You can filter the features by setting a threshold for the minimum activation rate.



Overview Feature Table Feature Detail

Ranking: Confusion Matrix Predicted label: Any True label: Different ⓘ

Order: ○ Ascending ● Descending Min. activation rate: 0.1 %

ID	ERR. RATE	ACT. RATE	ACT. DISTRIBUTION	TOP CLASS PROBABILITIES	EXAMPLE
2085	44.77%	0.869%		Bearish Neutral Bullish	i'll be in **at** 🤓😜</s><pad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

At the bottom of the tab, you can change the page of the table.

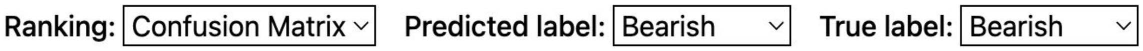
Feature Ranking

Three ways to rank the features:

- ID
    - Index of the feature in the SAE.
    - This is basically a random order.
- Activation rate
    - How often the feature activates.
- Confusion matrix
    - The model’s predictions and errors on the instances that activate the feature.

Confusion Matrix Ranking

- We calculate a confusion matrix for each feature based on the instances that cause that feature to activate.
- We can rank the features based on values derived from these confusion matrices.
- For this ranking, you choose both a predicted label and a true label.
    - This determines the value that is used to rank the features.

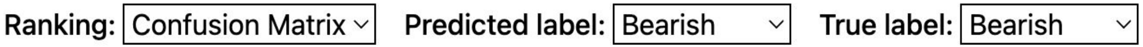


Ranking: Confusion Matrix

Predicted label: Bearish

True label: Bearish

Confusion Matrix Ranking - Example 1

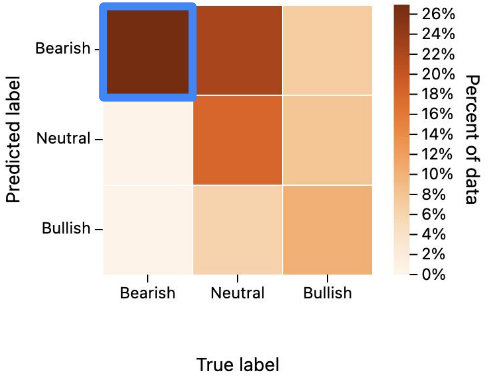


Ranking: Confusion Matrix

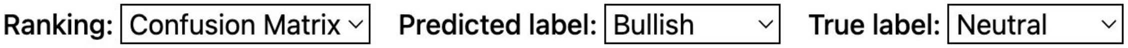
Predicted label: Bearish

True label: Bearish

- Predicted label is Bearish.
- True label is Bearish.
- This ranks the features by the percentage of instances that activate the feature where the model correctly predicts Bearish.



Confusion Matrix Ranking - Example 2

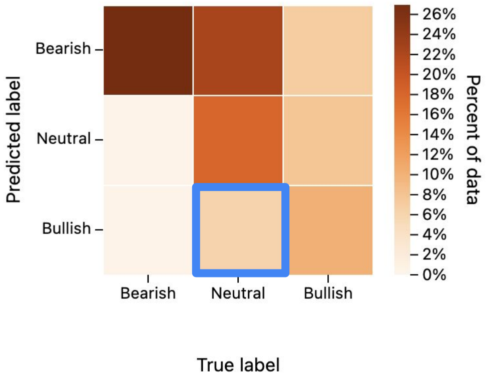


Ranking: Confusion Matrix

Predicted label: Bullish

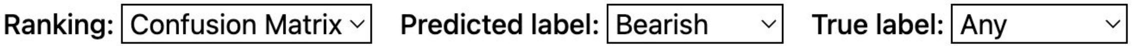
True label: Neutral

- Predicted label is Bullish.
- True label is Neutral.
- This ranks the features by the percentage of instances that activate the feature where the model predicts Bullish, but the true label is Neutral.



Confusion Matrix Ranking

- In addition to selecting the name of a class, you can also select the wildcard value “Any”.
- This represents any of the classes.

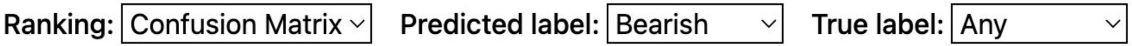


Ranking: Confusion Matrix

Predicted label: Bearish

True label: Any

Confusion Matrix Ranking - Example 3

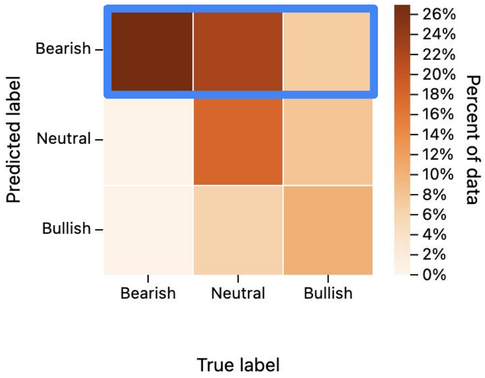


Ranking: Confusion Matrix

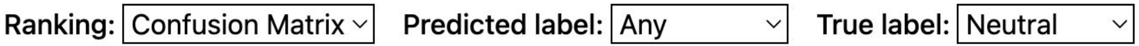
Predicted label: Bearish

True label: Any

- Predicted label is Bearish.
- True label is Any.
- This ranks the features by the percentage of instances that activate the feature where the model predicts Bearish.

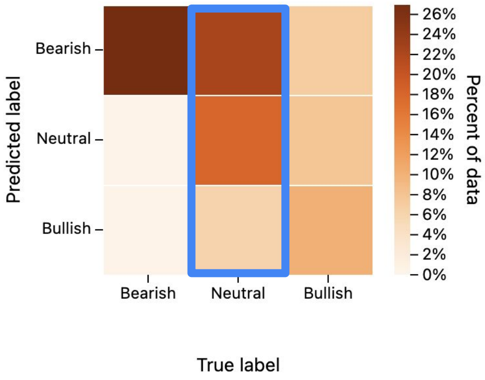


Confusion Matrix Ranking - Example 4



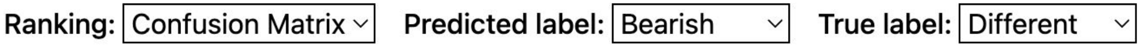
Ranking: Confusion Matrix ⌄ Predicted label: Any ⌄ True label: Neutral ⌄

- Predicted label is Any.
- True label is Neutral.
- This ranks the features by the percentage of instances that activate the feature where the true label is Neutral.



Confusion Matrix Ranking

- The second wildcard value is “Different”.
- This represents any class except the other one that is selected.
- For example:
  - Predicted label is Bearish
  - True label is Different
    - Different represents Bullish or Neutral.

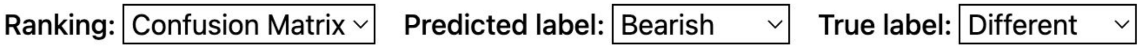


Ranking: Confusion Matrix

Predicted label: Bearish

True label: Different

Confusion Matrix Ranking - Example 5

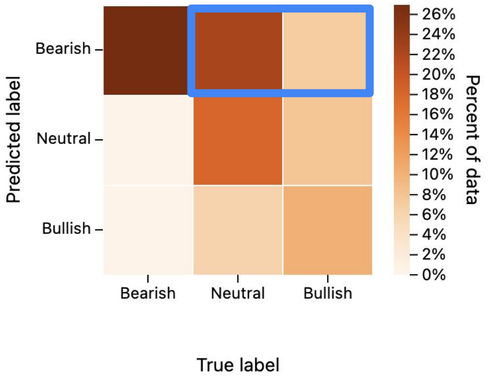


Ranking: Confusion Matrix

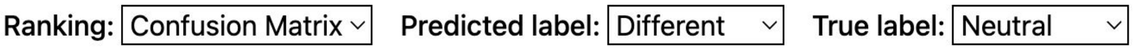
Predicted label: Bearish

True label: Different

- Predicted label is Bearish.
- True label is Different.
- This ranks the features by the percentage of instances that activate the feature where the model incorrectly predicts the label to be Bearish.



Confusion Matrix Ranking - Example 6

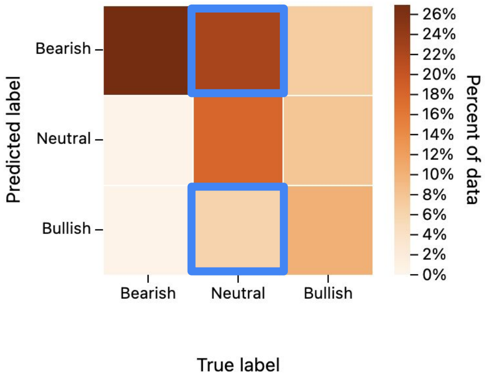


Ranking: Confusion Matrix

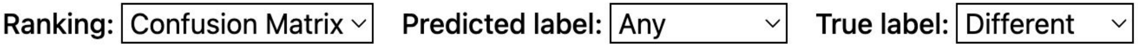
Predicted label: Different

True label: Neutral

- Predicted label is Different.
- True label is Neutral.
- This ranks the features by the percentage of instances that activate the feature where the true label is Neutral and the model predicts something else.

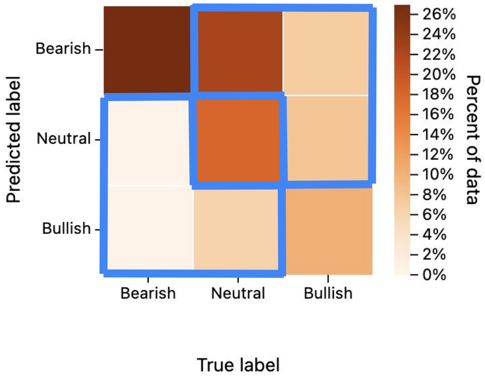


Confusion Matrix Ranking - Example 7

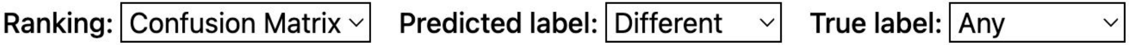


Ranking: Confusion Matrix Predicted label: Any True label: Different

- Predicted label is Any.
- True label is Different.
- This ranks the features by the percentage of instances that activate the feature where the model is incorrect.

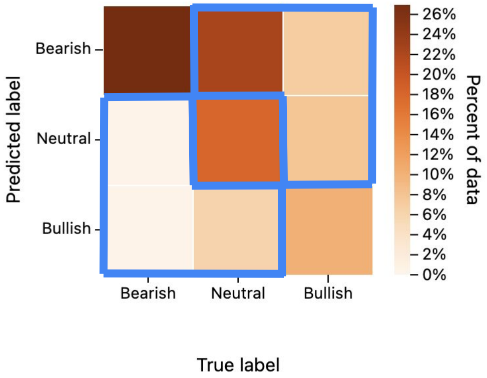


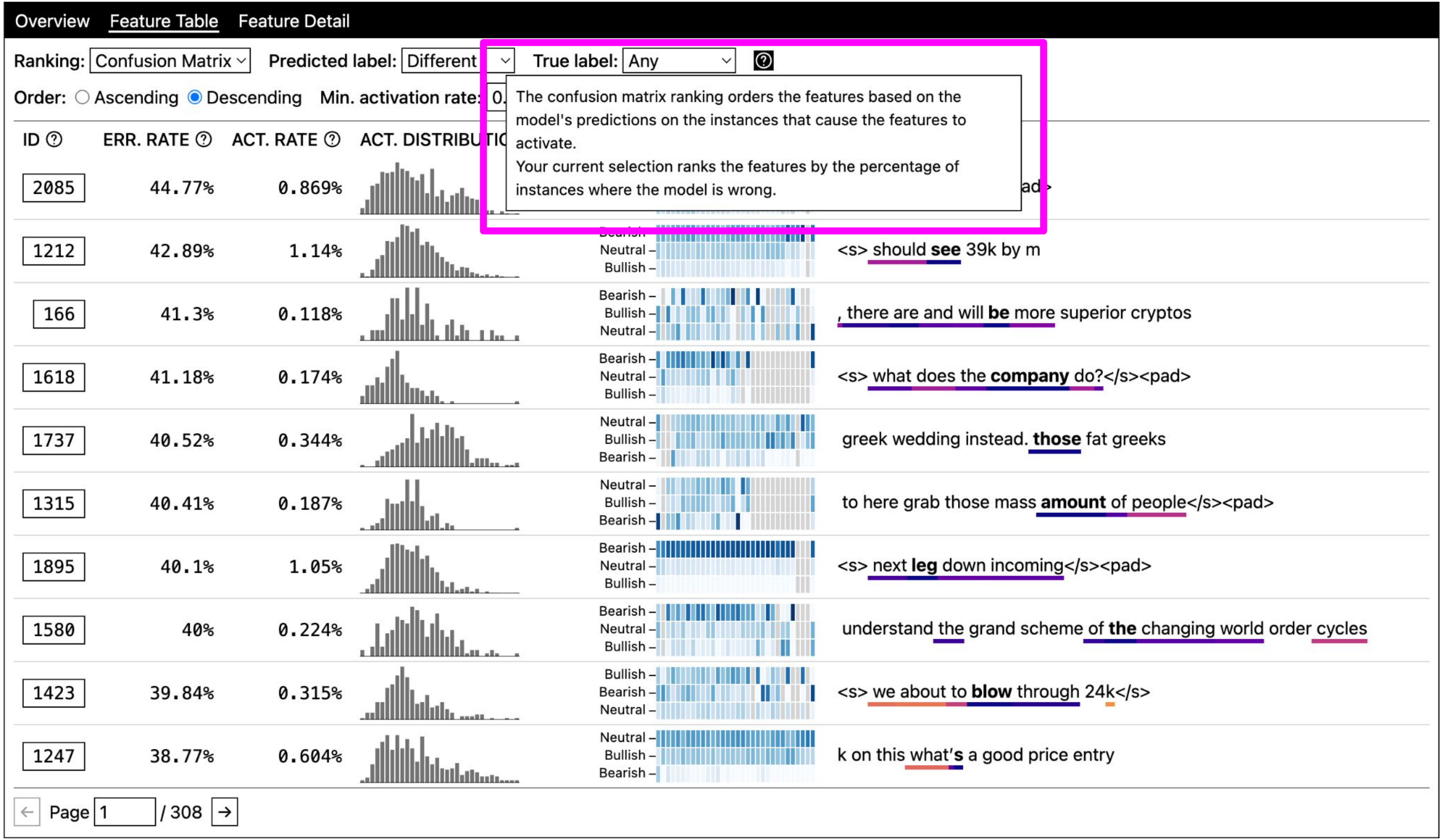
Confusion Matrix Ranking - Example 8



Ranking: Confusion Matrix ⌄ Predicted label: Different ⌄ True label: Any ⌄

- Predicted label is Different.
- True label is Any.
- This ranks the features by the percentage of instances that activate the feature where the model is incorrect. Same as previous case.





Ranking: Confusion Matrix Predicted label: Different True label: Any

Order: ○ Ascending ● Descending Min. activation rate: 0.

The confusion matrix ranking orders the features based on the model's predictions on the instances that cause the features to activate.
Your current selection ranks the features by the percentage of instances where the model is wrong.

ID	ERR. RATE	ACT. RATE	ACT. DISTRIBUTI		
2085	44.77%	0.869%			ad>
1212	42.89%	1.14%		Bearish Neutral Bullish	<s> should **see** 39k by m
166	41.3%	0.118%		Bearish Bullish Neutral	, there are and will **be** more superior cryptos
1618	41.18%	0.174%		Bearish Neutral Bullish	<s> what does the **company** do?</s><pad>
1737	40.52%	0.344%		Neutral Bullish Bearish	greek wedding instead. **those** fat greeks
1315	40.41%	0.187%		Neutral Bullish Bearish	to here grab those mass **amount** of people</s><pad>
1895	40.1%	1.05%		Bearish Neutral Bullish	<s> next **leg** down incoming</s><pad>
1580	40%	0.224%		Bearish Neutral Bullish	understand the grand scheme of **the** changing world order cycles
1423	39.84%	0.315%		Bullish Bearish Neutral	<s> we about to **blow** through 24k</s>
1247	38.77%	0.604%		Neutral Bullish Bearish	k on this what'**s** a good price entry

← Page 1 / 308 →

The question mark icon next to the ranking controls explains how you are currently ranking the features.

Interview

- Based on your exploration, do you have any ideas for how you might improve the model?

- How did SAEfarer support your model analysis?
- Did it help you to understand the behavior of the model?

- In what ways did it not support your model analysis?
- Were there any questions that you were unable to answer?
- Were there any tasks you were unable to perform?

- Were the visualizations useful?
- Were any of them unclear?

- Were the confusion matrix rankings helpful?
- Was it clear how to use it?
- How did you prefer to use the confusion matrix rankings?

- Are there any additional ways that would be helpful to rank the features?
- Are there any additional ways that would be helpful to filter the features?

- What aspects of SAEfarer did you like the most?
- What parts of the tool were the most helpful?

- What are SAEfarer’s biggest weaknesses or limitations?
- Are there any improvements or additional capabilities that you would want SAEfarer to have?
- Is there additional information that would be helpful to provide in the tool?

- Do you have any other feedback about SAEfarer?

Conclusion

- Thank you!
- We will send you your compensation soon.