Explainable Sentiment Analysis

GoogleResearchProgram@PERTH

By Violet Visionaries

Dipali Anil

Jaya Motwani

Benyapa Insawang

VIOLET VISIONARIES

CONTENT

INTRODUCTION

Natural Language Processing

- A subfield of linguistics, computer science, and artificial intelligence concerned with **the interactions between machine and human language.**
- The result is a machine **capable of "understanding" the contents of documents**, including the contextual nuances of the language within them.

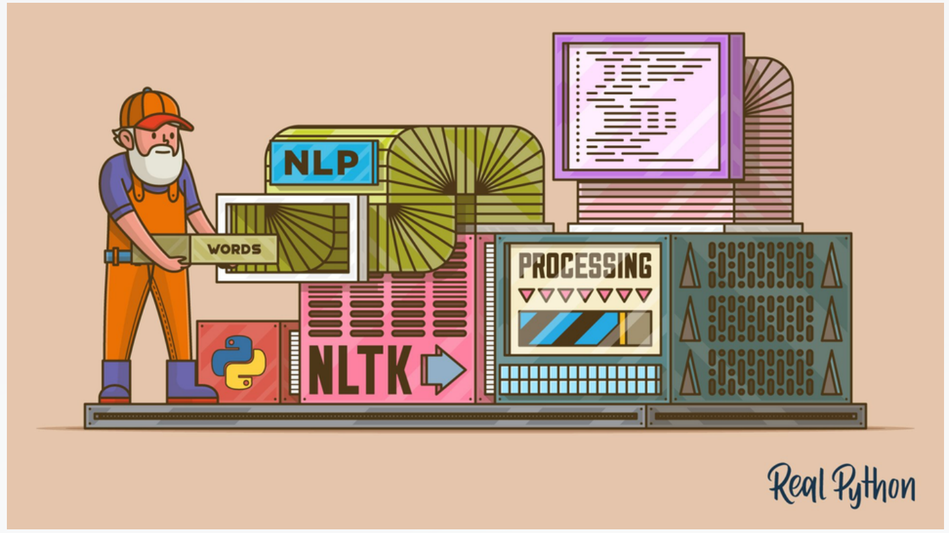


Fig. 1: Joanna Jablonski, https://realpython.com/nltk-nlp-python. Accessed 26 Aug. 2023.

INTRODUCTION

Text Classification

- To automatically classify the text documents into one or more defined categories. (label = classes or categories, data = text)

- Examples
  - Understanding audience sentiment from social media
  - Detection of spam and non-spam emails
  - Auto-tagging of customer queries
  - Categorization of news articles into defined topics

INTRODUCTION

Sentiment Analysis

The operation of understanding the intention or emotion behind a given piece of text.

★★★★★ **Can't say enough** 🥰

Reviewed in the United States on 12 August 2023

Size: Small | Colour: White-diva Blue Tie Dye | **Verified Purchase**

These are high quality at an affordable price. Literally I'm so shocked how well these fit. Better than big name brands. I feel the other bigger brands are made for taller ppl with longer torso. The waist band sit right at the center of my waist. Not too tight. SUPER flattering on the backside. Everything is perfect about these. Buying in like 5 more colors rn. Shorter girls must buy. I'm 5'3" 115lbs 24 inch waist 37 inch hips natural body it's hard to find shorts that look this good on.

One person found this helpful

POSITIVE

NEGATIVE

★★★☆☆ **Love the shorts but...**

Reviewed in Australia on 22 April 2023

Size: Large | Colour: Dark Black | **Verified Purchase**

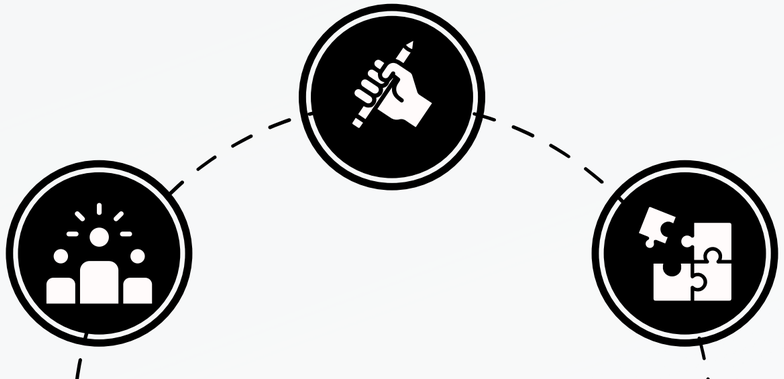
I look the fit of the shorts but I've worn them once and they've already pilled 😞 disappointing when you're paying $50+

RESEARCH QUESTION...?

- Why does the model make the decision like that?
- What features of the input does the model consider the most while making predictions?

WHY EXPLANATION IS IMPORTANT?

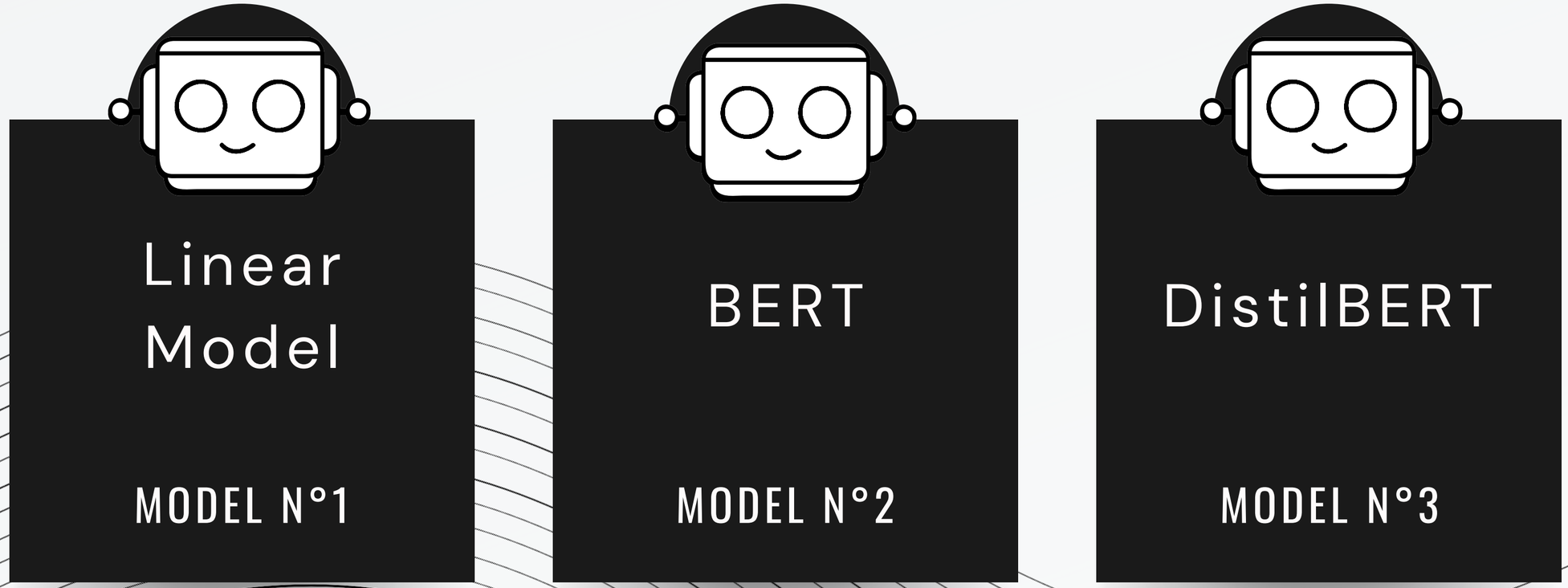
- Why is explanation important?
- Providing more information: Evidence
- Assisting in justifying the model's performance
- Establishing trust in the model



RESEARCH AIM

Explain the model's prediction results
in the sentiment analysis task

METHODOLOGY



METHODOLOGY

	Linear Model	BERT Model	DistilBERT Model
Architecture	Sequential Neural Network	Transformer	Transformer (less layers than BERT)
Parameters	100K	Base – 110M Large – 340M	66M
Speed	Fast	Slow	Faster than BERT
Accuracy	Low	High	High (97% of BERT)
Pre-trained	No	Yes	Yes
Positional encoding	No	Yes	Yes

METHODOLOGY

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

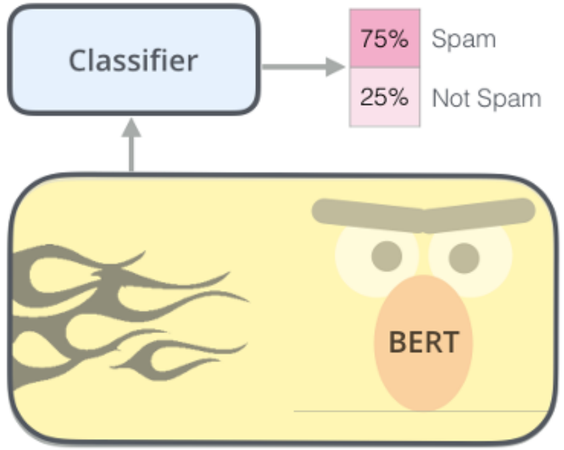
Model:

Dataset:

Objective: Predict the masked word (langauge modeling)

2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step



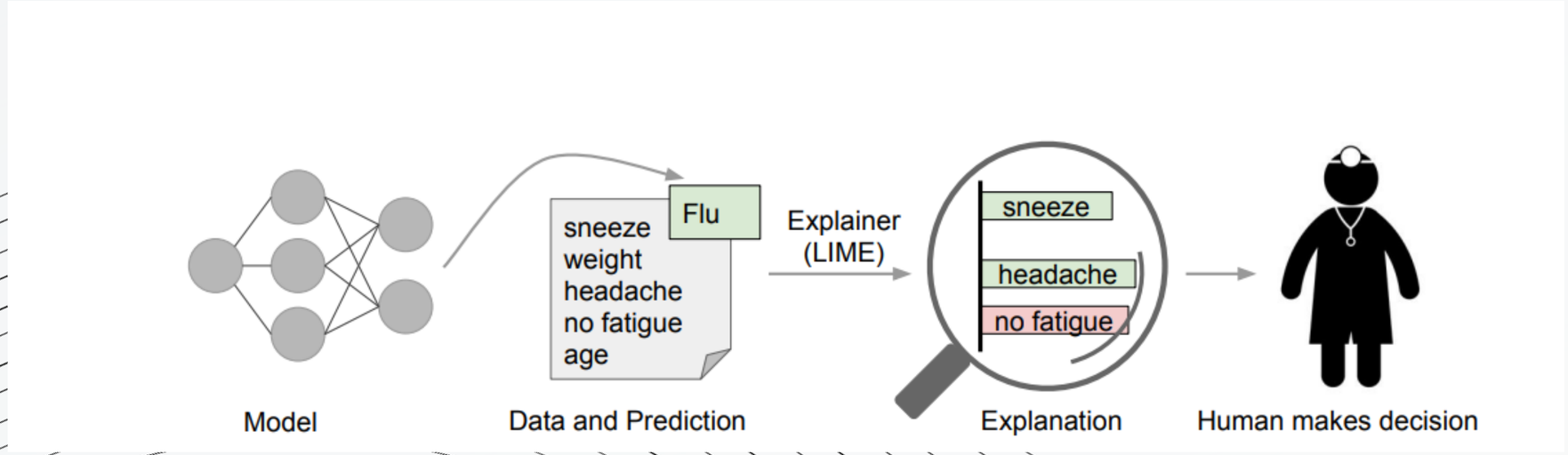
Model: (pre-trained in step #1)

Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

METHODOLOGY

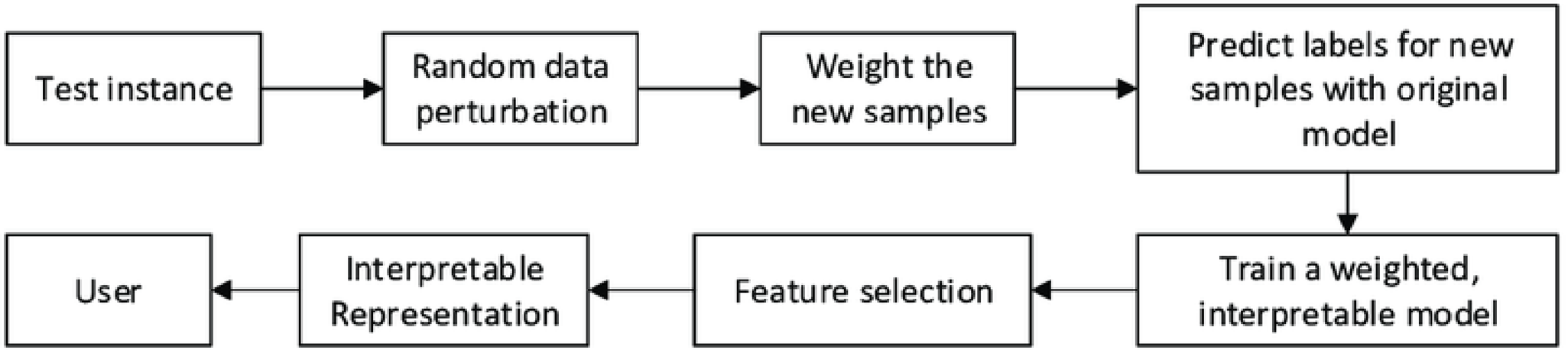
HOW TO INTERPRET THE MODEL PREDICTION...?



METHODOLOGY

HOW LIME WORKS....

LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANANTIONS



DATASET: INTRODUCTION

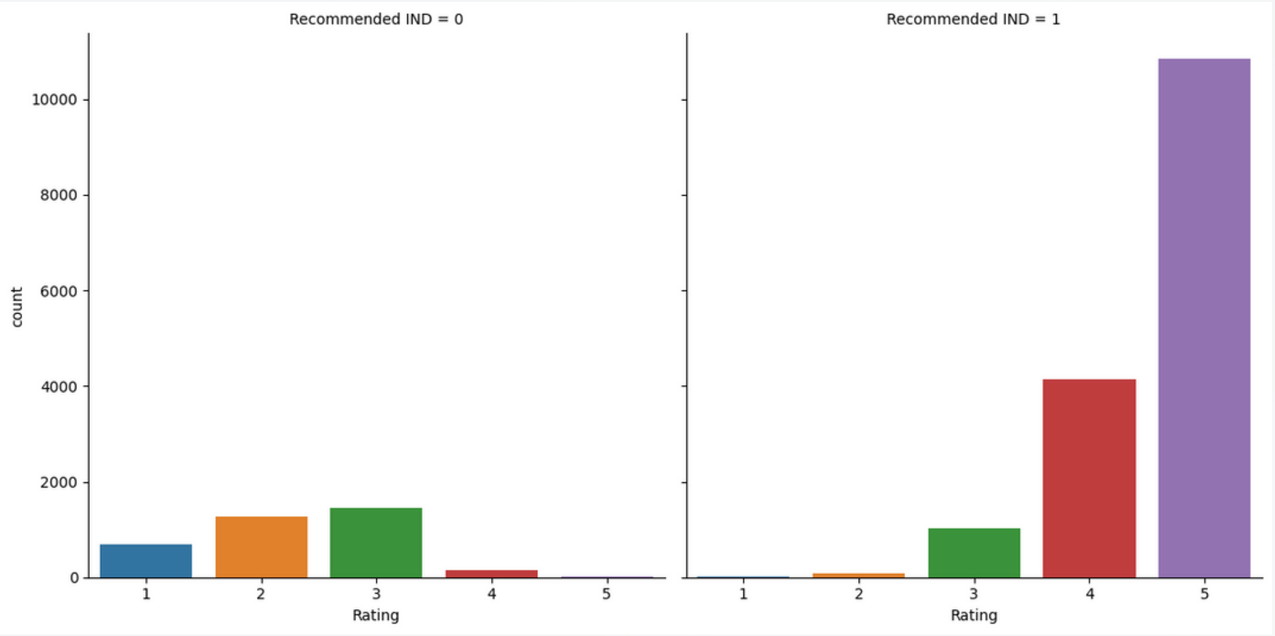
- Womens Clothing E-Commerce Reviews
- 23,486 observations of 11 variables
- Train/Test split: 80/20
- Target: **“Recommended”** or **“Not Recommended”**

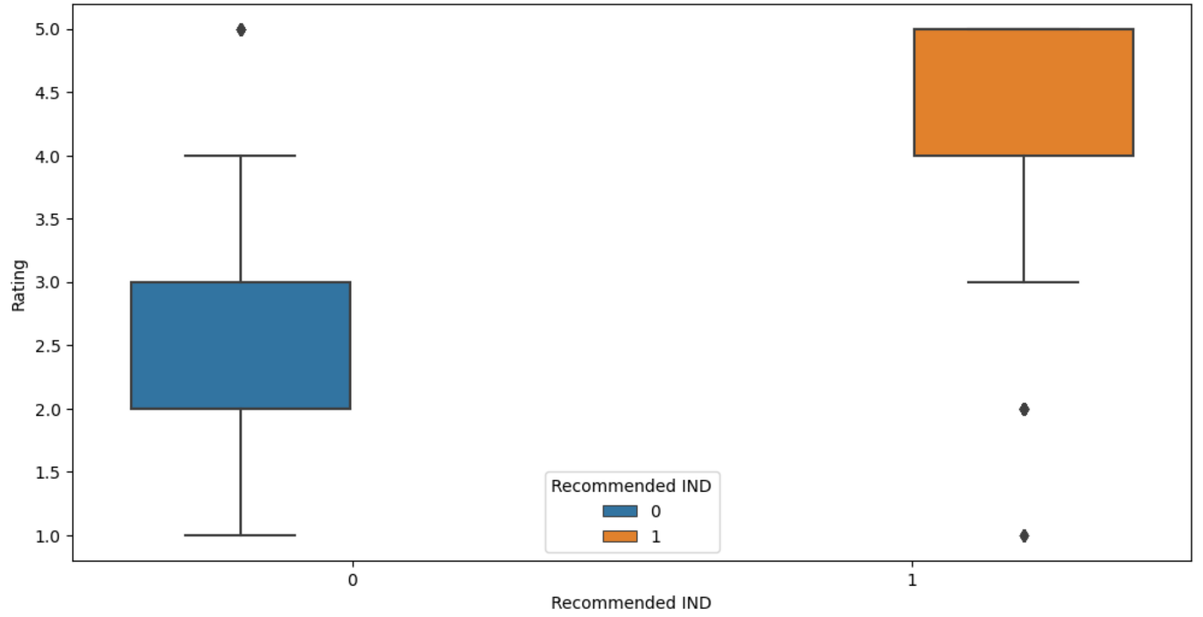
DATASET: PRE-PROCESSING

- Pre-processing:
  - Concatenation “Title” and “Review Text”
  - Drop NAs
  - Preproc method: Ktrain library

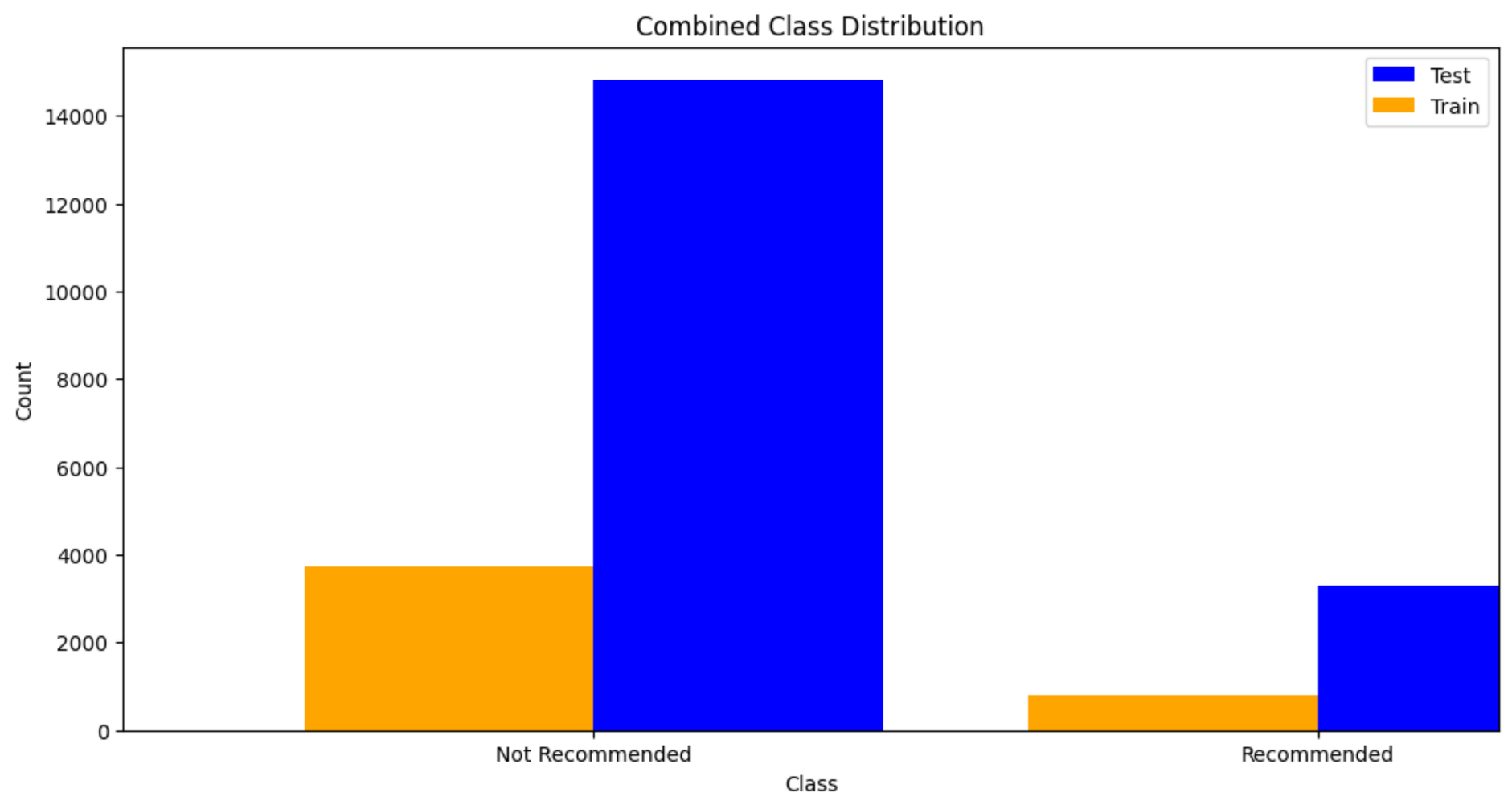
Unnamed: 0	Clothing ID	Age	Title	Review Text	Rating	Recommended IND	Positive Feedback Count	Division Name	Department Name	Class Name	Concat
21331	1092	31	Versatile	I passed over this dress online thinking i'd n...	5	1	0	General	Dresses	Dresses	Versatile I passed over this dress online thin...
8113	406	33	Such a unique, flattering swimsuit!	I am so excited about this swimsuit. i have ne...	5	1	0	Initmates	Intimate	Swim	Such a unique, flattering swimsuit! I am so ex...
23036	1008	31	The perfect skirt for all year, every year.	To start: i'm 135 lbs, 5'7", 34c, 28 jeans, si...	5	1	0	General Petite	Bottoms	Skirts	The perfect skirt for all year, every year. To...
316	836	59	Love this blouse	I really like this blouse a lot. very very eas...	5	1	0	General	Tops	Blouses	Love this blouse I really like this blouse a l...
6781	1030	39	So crazy comfortable	I haven't worn jeans in a year because i felt ...	5	1	0	General	Bottoms	Jeans	So crazy comfortable I haven't worn jeans in a...

DATASET: VISUALISATION





DATASET: VISUALISATION



CASE STUDY

Explanation of 3 models

```
print("Linear model")
l_predictor.explain(review)
```

Linear model
y=Recommended IND (probability **0.925**, score **2.517**) top features

Contribution[?]	Feature
+2.046	Highlighted in text (sum)
+0.471	<BIAS>

great spring sweater this sweater is classy and comfortable. it has an underlay which makes it truly unique.

```
print("BERT model")
b_predictor.explain(review)
```

BERT model
y=Recommended (probability **1.000**, score **10.398**) top features

Contribution[?]	Feature
+10.080	Highlighted in text (sum)
+0.318	<BIAS>

great spring sweater this sweater is classy and comfortable. it has an underlay which makes it truly unique.

```
print("DistilBERT model")
d_predictor.explain(review)
```

DistilBERT model
y=1 (probability **1.000**, score **11.185**) top features

Contribution[?]	Feature
+10.981	Highlighted in text (sum)
+0.204	<BIAS>

great spring sweater this sweater is classy and comfortable. it has an underlay which makes it truly unique.

DEMO

CASE STUDY

Correct prediction
with related-important words

```
print("DistilBERT model")
d_predictor.explain(review)
```

DistilBERT model
y=1 (probability **1.000**, score **7.799**) top features

Contribution?	Feature
+7.826	Highlighted in text (sum)
-0.027	<BIAS>

super cute and comfy pull over. sizing is accurate. material has a little bit of stretch.

CASE STUDY

Correct prediction
with nonrelated-important words

```
print("DistilBERT model")
d_predictor.explain(review)
```

DistilBERT model
y=1 (probability **0.997**, score **5.688**) top features

Contribution?	Feature
+5.478	Highlighted in text (sum)
+0.210	<BIAS>

cream coloredthe top is more of a cream color than an ivory. it's also a bit boxer than anticipated. fit is true to size.

CASE STUDY

Incorrect prediction with related-important words

```
print("DistilBERT model")
d_predictor.explain(review)
```

DistilBERT model
y=1 (probability **0.724**, score **0.966**) top features

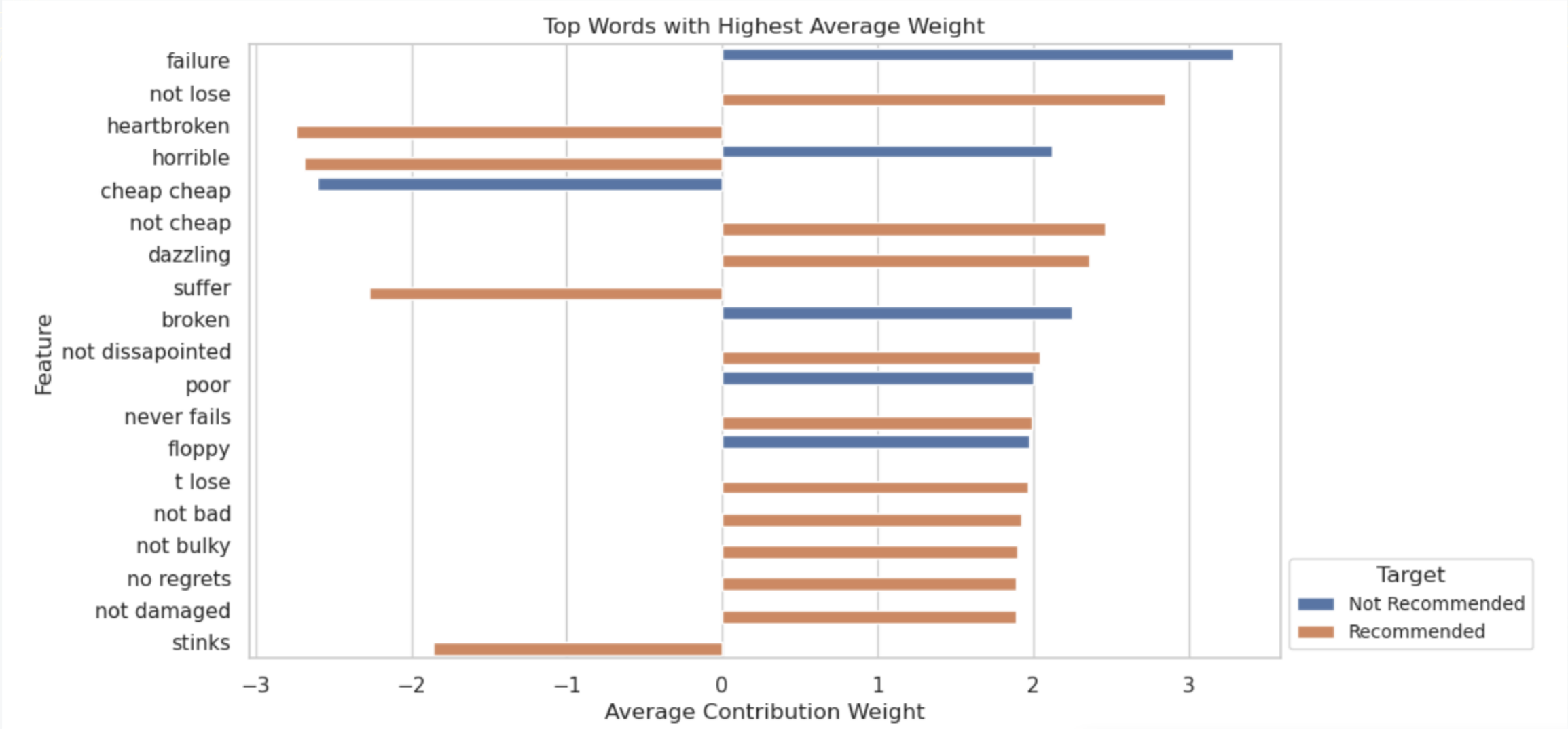
Contribution[?]	Feature
+0.963	Highlighted in text (sum)
+0.003	<BIAS>

gorgeous, but...i bought a 2p when i should've bought a 0p. i was swimming in the 2p. however, it didn't matter, because everything was itchy, and i'm not even allergic to wool. the beautiful buttons are oddly placed to achieve that cool design, and just like other reviewers commented, you have to twist and contort them to get them to button...which, inevitably, will lead to them popping off. and i didn't see an extra one in case that happened. the coat is stunning, there's no argument about that, and i welcome

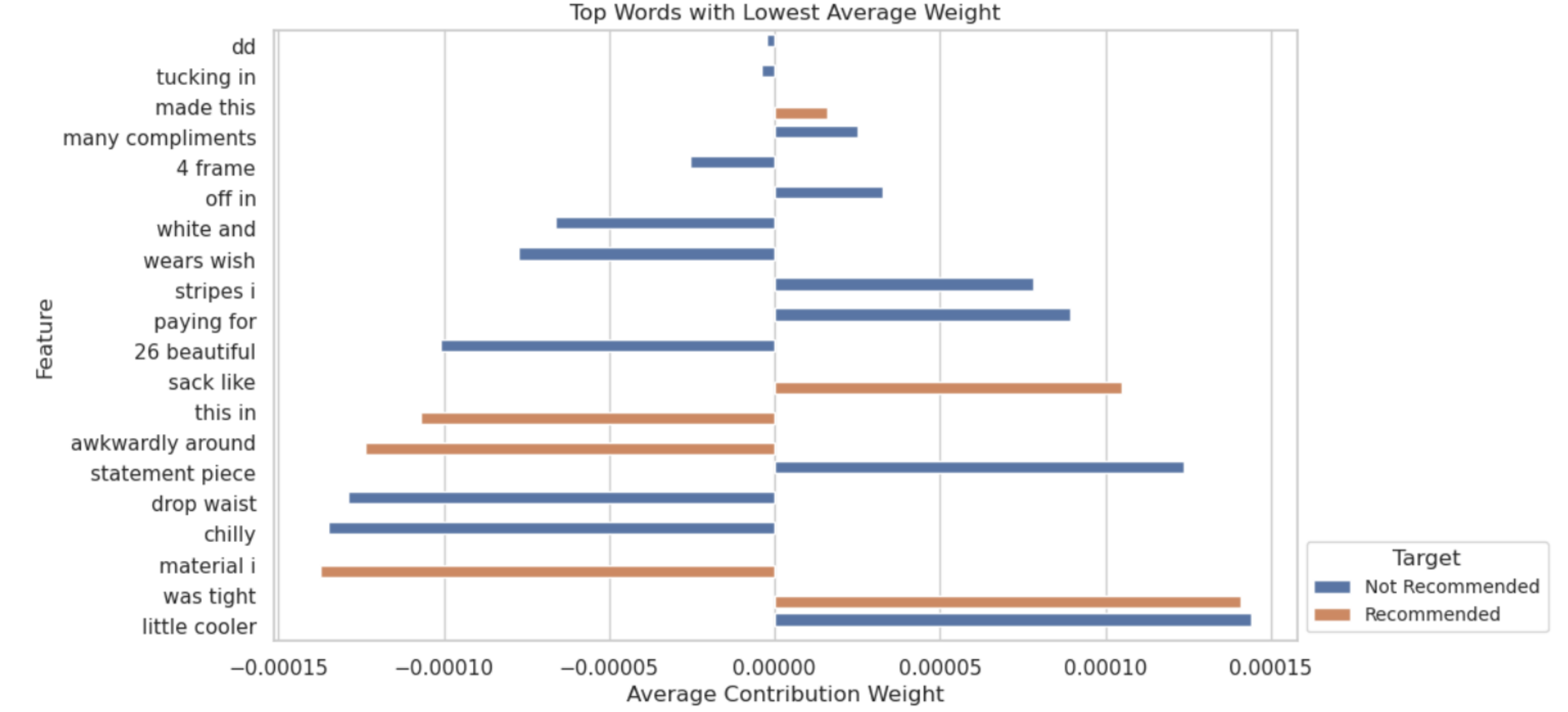
EVALUATION

	Linear model	BERT Model	DistilBERT Model
Accuracy	0.89	0.93	0.93
Precision	NR – 0.79 R – 0.9 WA – 0.88	NR – 0.79 R – 0.97 WA – 0.93	NR – 0.79 R – 0.96 WA – 0.93
Recall	NR – 0.51 R – 0.97 WA – 0.89	NR – 0.84 R – 0.95 WA – 0.93	NR – 0.83 R – 0.95 WA – 0.93
F1 score	NR – 0.62 R – 0.93 WA – 0.88	NR – 0.82 R – 0.96 WA – 0.93	NR – 0.81 R – 0.96 WA – 0.93

EVALUATION


Top Words with Highest Average Weight
failure
not lose
heartbroken
horrible
cheap cheap
not cheap
dazzling
suffer
broken
not dissapointed
poor
never fails
floppy
t lose
not bad
not bulky
no regrets
not damaged
stinks
Feature
−3
−2
−1
0
1
2
3
Average Contribution Weight
Target
Not Recommended
Recommended

EVALUATION


Top Words with Lowest Average Weight
dd
tucking in
made this
many compliments
4 frame
off in
white and
wears wish
stripes i
paying for
26 beautiful
sack like
this in
awkwardly around
statement piece
drop waist
chilly
material i
was tight
little cooler
Feature
−0.00015
−0.00010
−0.00005
0.00000
0.00005
0.00010
0.00015
Average Contribution Weight
Target
Not Recommended
Recommended

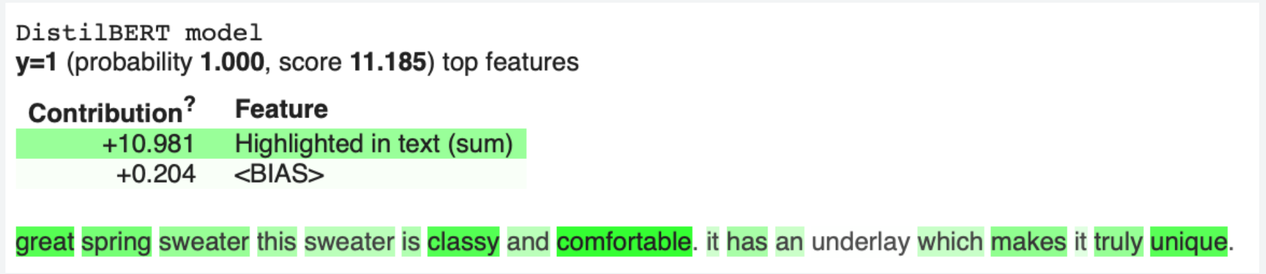
INSIGHTS

- Classification might not always be a true reflection of sentiment expressed through Review.
- The criteria that AI models prioritize as significant for classification might not align with what humans would deem as important.
- The actual interpretation still requires people.

INSIGHTS

- "The review was likely classified as recommended because the text has an overall positive tone, indicating that the reviewer felt positively about the product (and therefore likely enjoyed it and feels that other people would enjoy it as well.) Most people would only recommend something if they enjoyed it."

- "The reviewer recommended the product because it is comfortable. It is more important that it is comfortable than how classy it is or the underlay, because you won't wear something if it is not comfortable."


DistilBERT model
y=1 (probability 1.000, score 11.185) top features

Contribution?	Feature
+10.981	Highlighted in text (sum)
+0.204	<BIAS>

great spring sweater this sweater is classy and comfortable. it has an underlay which makes it truly unique.

LIMITATIONS

- Since LIME is based on the underlying black box model to generate new pairing of data sets (permuted reviews and predicted target), the explanation is not always correct.
- It is also subjective and requires human interpretation.
- LIME explanations can be manipulated by the data scientist to hide biases.
- The method is still in development phase and many problems need to be solved before it can be safely applied.
- LIME generates local explanations.

FUTURE SCOPE

- In this particular example, more features e.g., Ratings can be used as an input in black box model for better prediction and interpretation.
- Alternatives to LIME:
  - Attention Weights
  - SHAP (SHapley Additive exPlanations)
- Automation of Interpretation

Q&A
THANK YOU