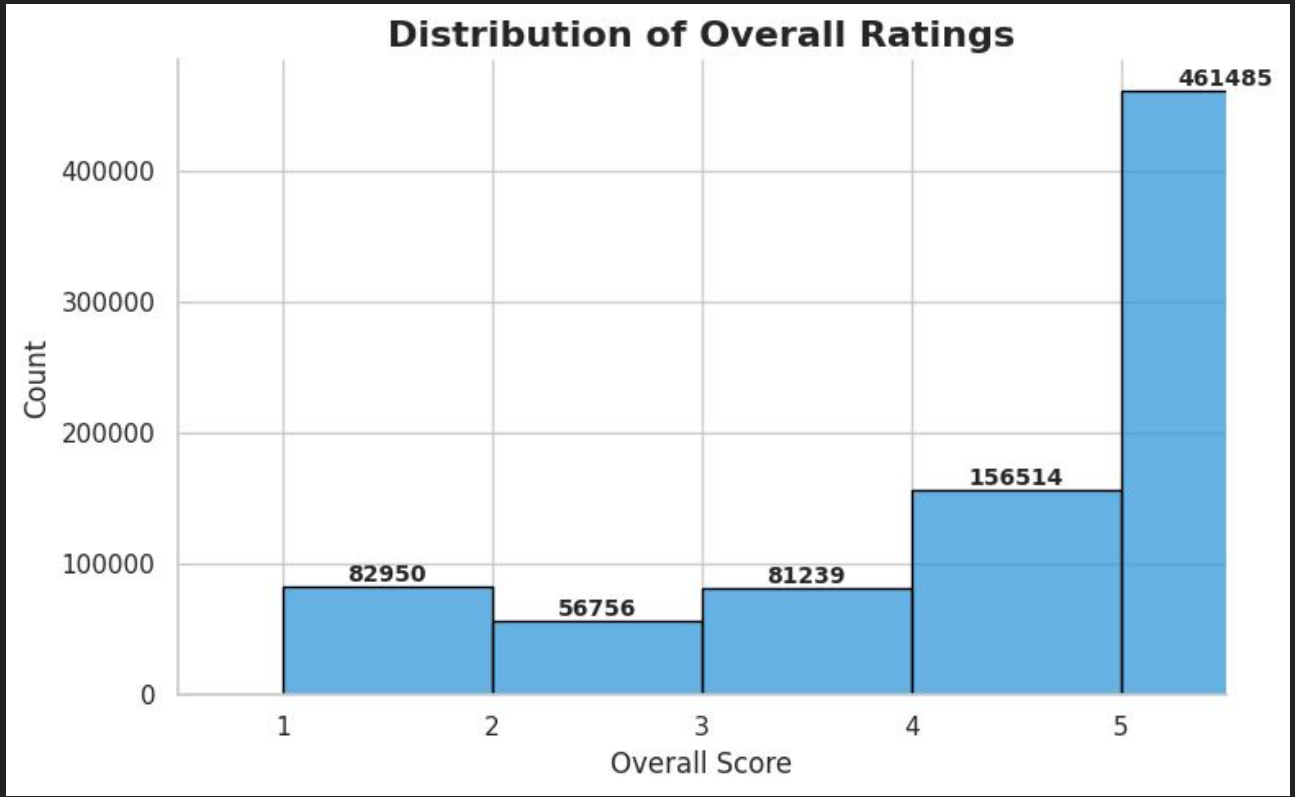
Amazon reviews sentiment analysis

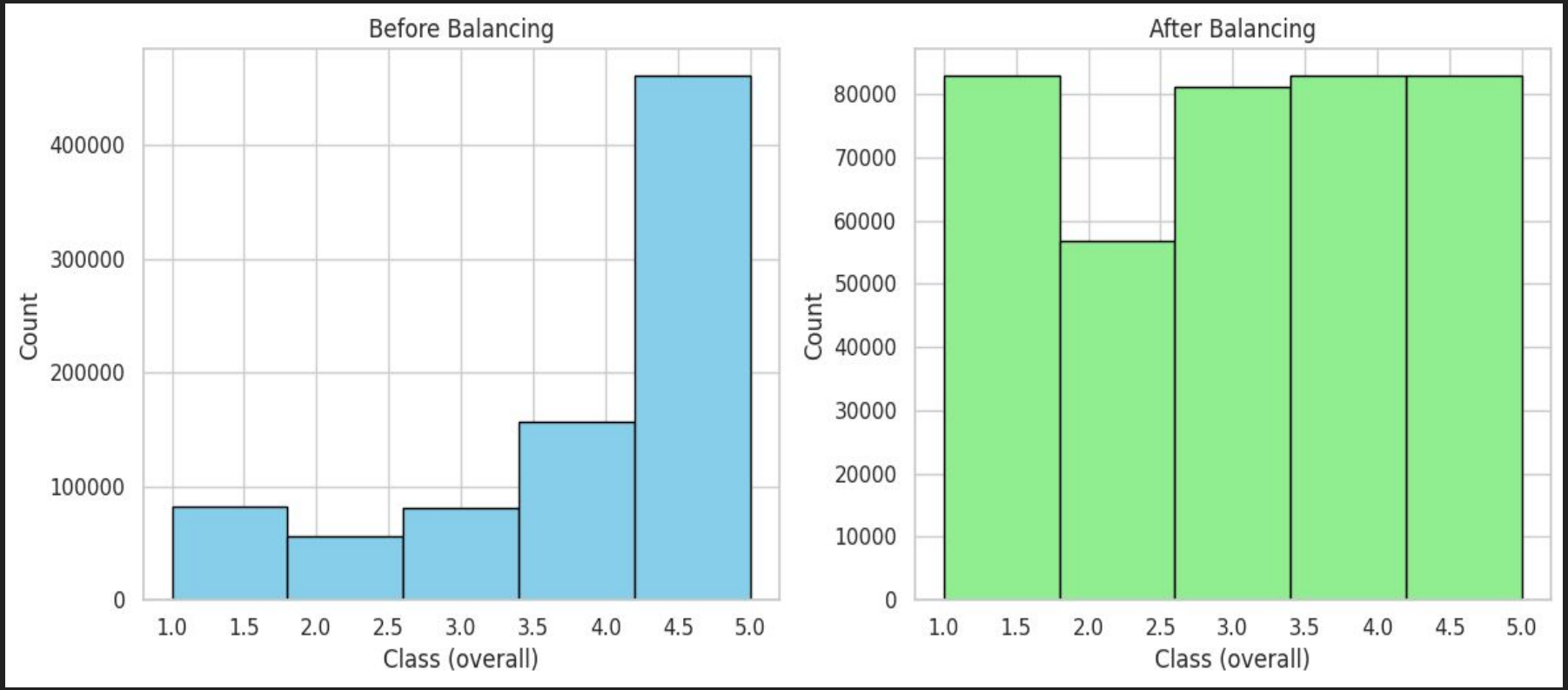
EDA

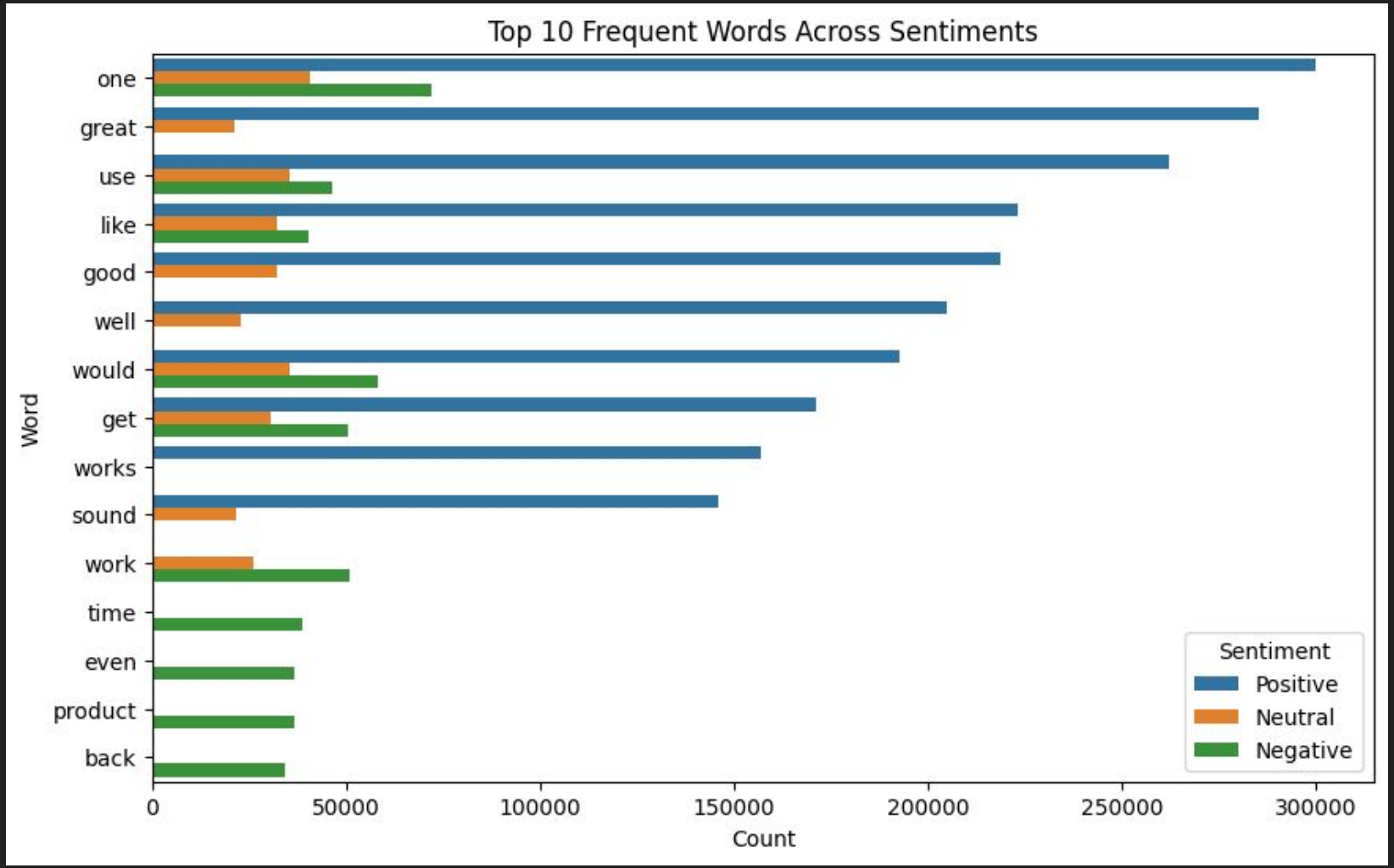

Distribution of Overall Ratings
461485
400000
300000
Count
200000
156514
100000
82950
81239
56756
0
1
2
3
4
5
Overall Score

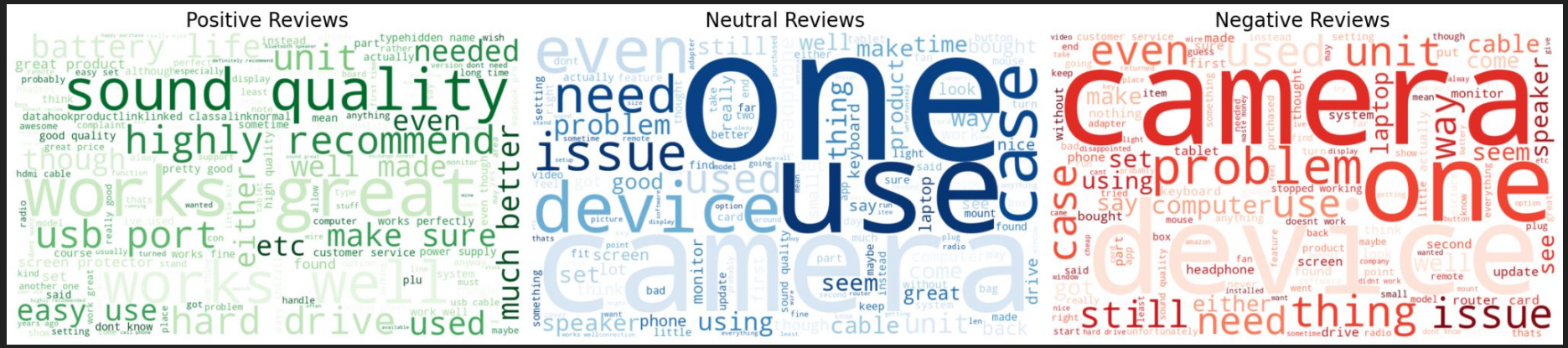
Balancing the Dataset using Random Under-Sampling

In this section, we balance the dataset by reducing the number of samples in overrepresented classes using a mild under-sampling strategy. The goal is to make the target variable (overall) more evenly distributed without losing too much information.

- Original class distribution: Counter({5: 461485, 4: 156514, 1: 82950, 3: 81239, 2: 56756})
- Target number of samples per class (median): 82950
- Mean number of samples: 167788
- Sampling strategy: {2: 56756, 5: 82950, 4: 82950, 3: 81239, 1: 82950}
- Class distribution after balancing: Counter({1: 82950, 4: 82950, 5: 82950, 3: 81239, 2: 56756})
- Final balanced DataFrame shape: (386845, 11)


Before Balancing
Count
400000
300000
200000
100000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Class (overall)
After Balancing
Count
80000
70000
60000
50000
40000
30000
20000
10000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Class (overall)

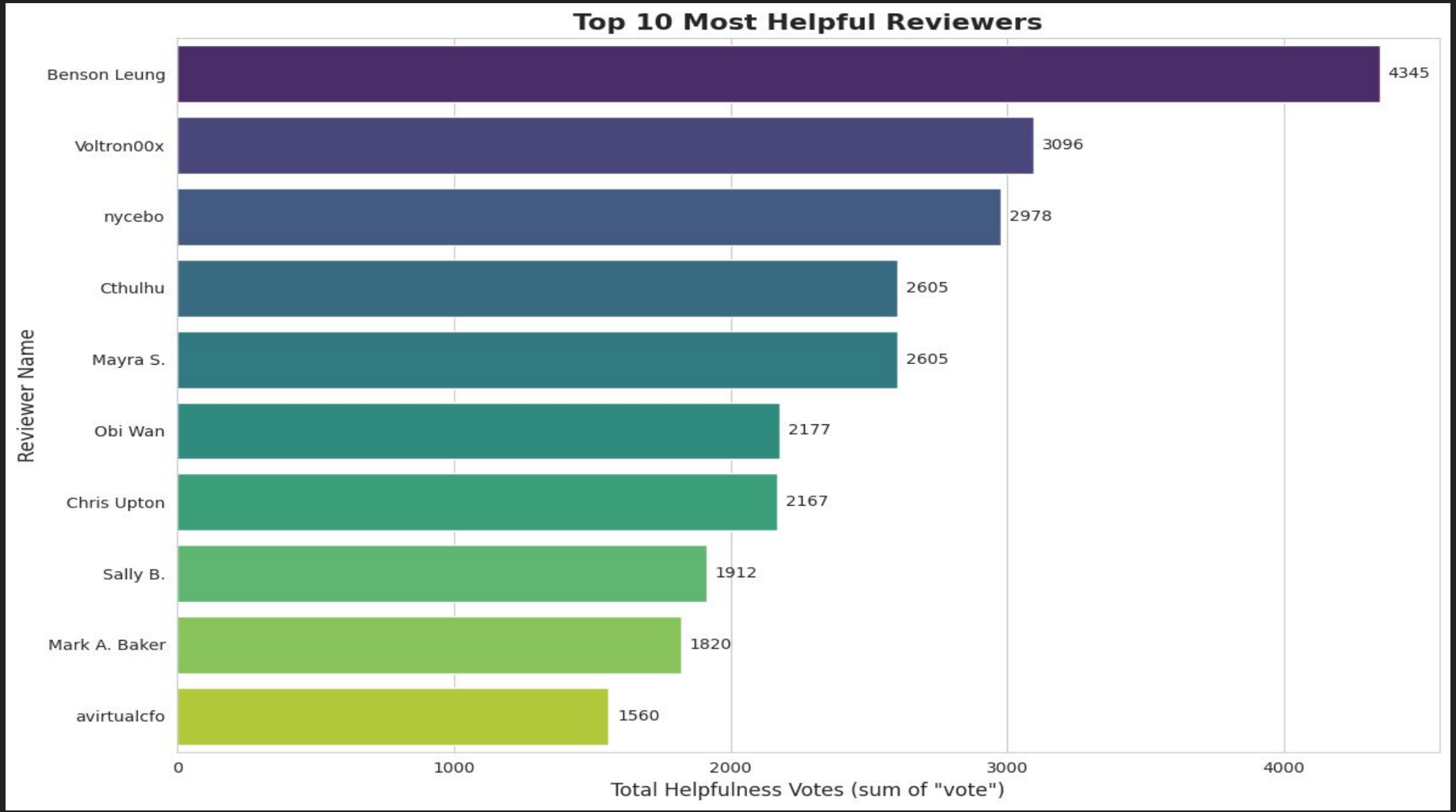

Top 10 Frequent Words Across Sentiments
one
great
use
like
good
well
would
get
works
sound
work
time
even
product
back
Word
0
50000
100000
150000
200000
250000
300000
Count
Sentiment
Positive
Neutral
Negative



The frequency and Word Cloud analysis revealed that all three sentiment classes share a substantial overlap in vocabulary — especially words like *use, work, good, one,* and *sound*.

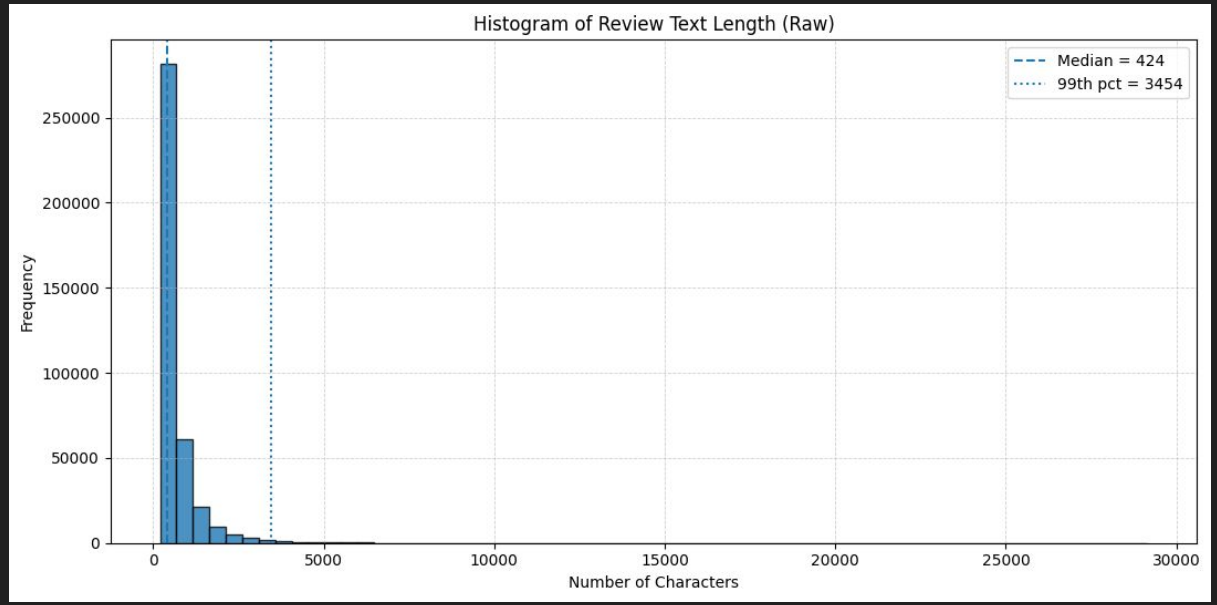
- **Positive reviews** highlight satisfaction through words like *works, great, recommend,* and *easy*.
- **Negative reviews** emphasize issues with *problem, broken, issue,* and *need*.
- **Neutral reviews** focus on descriptive, emotionless terms such as *camera, unit,* and *use*.

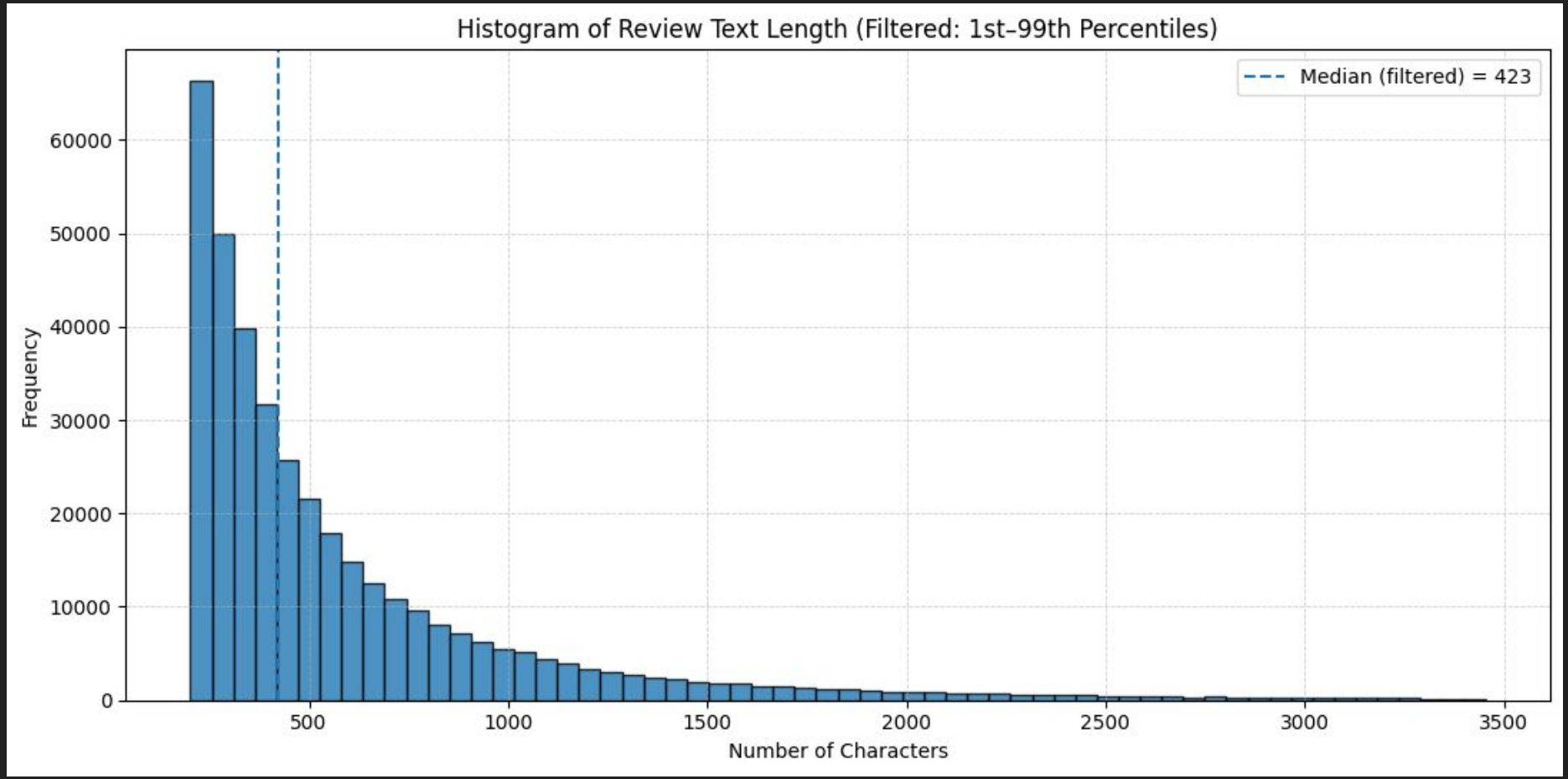
Overall, while word frequencies are similar, **context and phrasing** determine the true sentiment. Therefore, deeper models (like sentiment classifiers) should consider **phrase structure** and **contextual meaning**,
not just raw word counts.

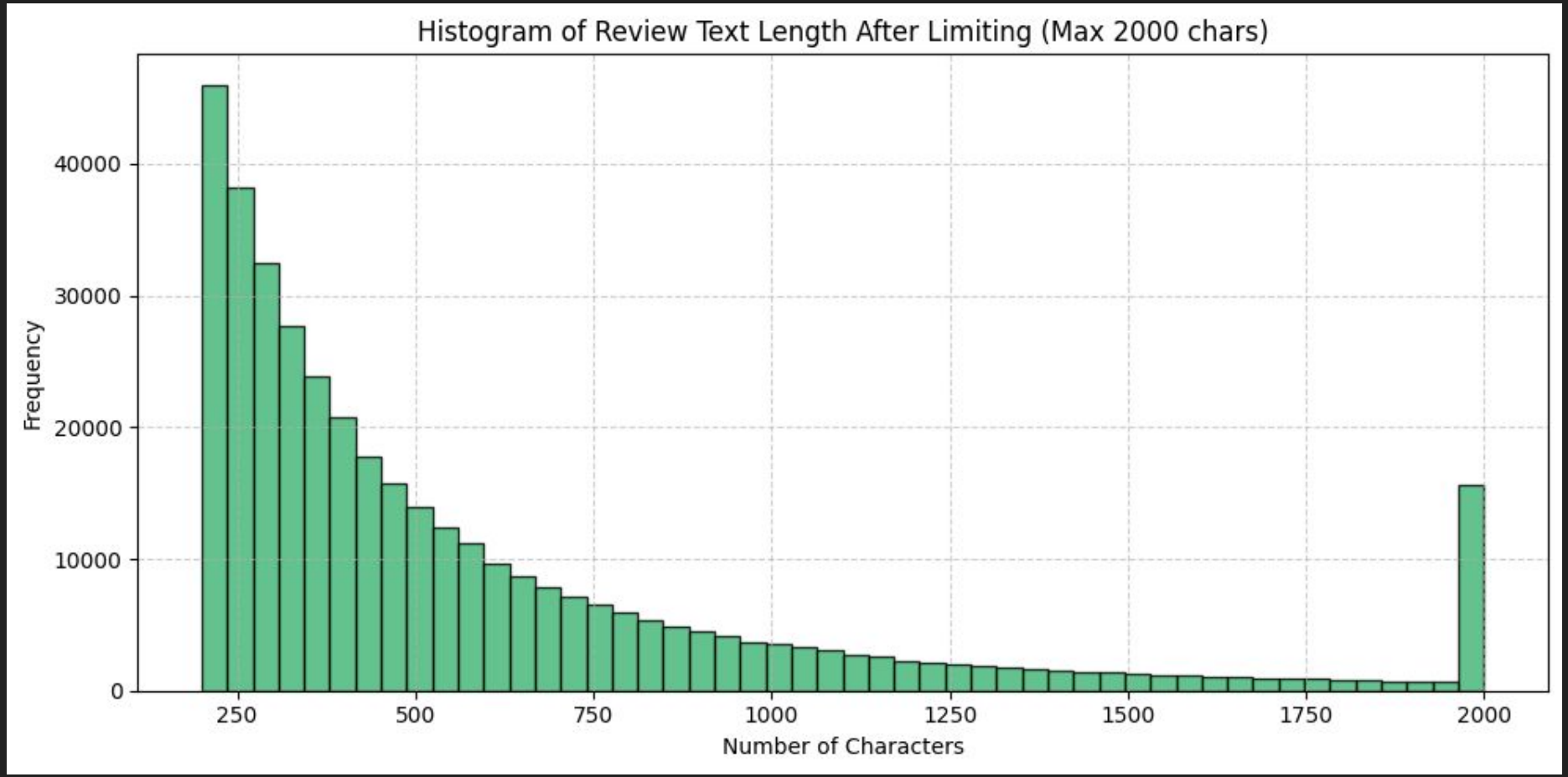

Top 10 Most Helpful Reviewers
Benson Leung
Voltron00x
nycebo
Cthulhu
Mayra S.
Obi Wan
Chris Upton
Sally B.
Mark A. Baker
avirtualcfo
4345
3096
2978
2605
2605
2177
2167
1912
1820
1560
Reviewer Name
0
1000
2000
3000
4000
Total Helpfulness Votes (sum of "vote")

Reviews length

- Min length: 200
- Max length: 29146
- Mean length: 642.49
- Median length: 424
- 1st percentile: 202
- 99th percentile: 3454




Histogram of Review Text Length (Filtered: 1st–99th Percentiles)
Median (filtered) = 423
Frequency
60000
50000
40000
30000
20000
10000
0
500
1000
1500
2000
2500
3000
3500
Number of Characters


Histogram of Review Text Length After Limiting (Max 2000 chars)
Frequency
0
10000
20000
30000
40000
250
500
750
1000
1250
1500
1750
2000
Number of Characters

Top 10 Products based on 5-Star Reviews Count

	brand	title	five_star_count
0	Cambridge Soundworks	OontZ Angle 3 Enhanced Stereo Edition IPX5 Splashproof Portable Bluetooth Speaker with Volume Booster AMP 10 Watts Power, Custom Bass Radiator, 100' Wireless Range Bluetooth 4.2	1163
1	NETGEAR	NETGEAR N300 WiFi Range Extender (EX2700)	659
2	Roku	Roku Streaming Stick (3600R) - HD Streaming Player with Quad-Core Processor	600
3	StarTech	StarTech USB 2.0 to SATA IDE Adapter (USB2SATAIDE)	598
4	Logitech	Logitech M570 Wireless Trackball Mouse – Ergonomic Design with Sculpted Right-hand Shape, Compatible with Apple Mac and Microsoft Windows Computers, USB Unifying Receiver, Dark Gray	552
5	Samsung	Samsung 850 EVO 500GB 2.5-Inch SATA III Internal SSD (MZ-75E500B/AM)	529
6	Asus	ASUS Tri-Band Gigabit (AC3200) WiFi Router (Up to 3167 Mbps) with MU-MIMO to ensure Lag-Free Gaming, AiProtection network security powered by Trend Micro, Adaptive QoS and Parental Control (RT-AC3200)	471
7	VideoSecu	VideoSecu ML531BE TV Wall Mount for Most 27"-55" LED LCD Plasma Flat Screen Monitor up to 88 lb VESA 400x400 with Full Motion Swivel Articulating 20 in Extension Arm, HDMI Cable & Bubble Level WP5	448
8	Arlo Technologies, Inc	Arlo - Wireless Home Security Camera System | Indoor/Outdoor | 2 camera kit (Discontinued)	433
9	Samsung	Samsung 850 EVO 250GB 2.5-Inch SATA III Internal SSD (MZ-75E250B/AM)	431

Average Score for Top 10 Brands

- *Anker (4.24)*
- *AmazonBasics (4.20)*
- *Sabrent (4.02)*
- *SanDisk (3.95)*
- *Logitech (3.95)*
- *Samsung (3.94)*
- *Sony (3.93)*
- *TP-LINK (3.92)*
- *Asus (3.85)*
- *NETGEAR (3.83)*

Amazon reviews sentiment analysis

Warranty Analyzing

Part 2

Problem: Analyzing customer sentiment specifically about warranty and customer support aspects from product reviews is challenging because warranty-related content is mixed with general product reviews, making it difficult to extract meaningful insights about warranty satisfaction.Challenge: Traditional sentiment analysis treats all reviews equally, but warranty sentiment requires identifying warranty-related reviews first, then analyzing their sentiment separately to provide actionable insights for product warranty policies.

Dataset columns:

- `fullText` → Combined review text and summary
- `overall` → Rating (1–5)
- `asin` → Product ID linked to brand/title metadata

Step 1

Semantic Similarity Filtering: Use Sentence-BERT to identify warranty-related reviews by calculating cosine similarity between review text and predefined warranty-related phrases, filtering reviews with similarity scores above 0.5 threshold.

Step 2

Review Embedding Generation: Encode all 386,845 reviews into 384-dimensional embeddings using the 'all-MiniLM-L6-v2' model, creating a semantic representation that captures warranty-related content meaning.

```
# Our target concept, described by a RICH set of phrases
query_phrases = [
    'product warranty and guarantee',
    'customer support for replacement or repair',
    'return policy and money back',
    'defective item and exchange process',
    'lifetime warranty',
    'limited warranty terms',
    'one-year warranty coverage',
    'helpful and responsive customer service',
    'contacting support was easy',
    'they never replied to my email',
    'long wait time for support',
    'the device stopped working after a week',
    'it was broken on arrival (DOA - Dead on Arrival)',
    'product malfunction and failure',
    'getting a full refund was straightforward',
    'return merchandise authorization (RMA) process'
]
```

step 3

Sentiment Analysis on Filtered Data: Apply sentiment analysis specifically to the filtered warranty-related reviews (4,627 reviews, 1.20% of total) using their original rating scores to calculate average warranty sentiment per product.

Step 4

Product-Level Aggregation and Ranking: Group results by product (ASIN), calculate average warranty sentiment scores, merge with product metadata, and rank products by warranty satisfaction to identify best and worst performers.

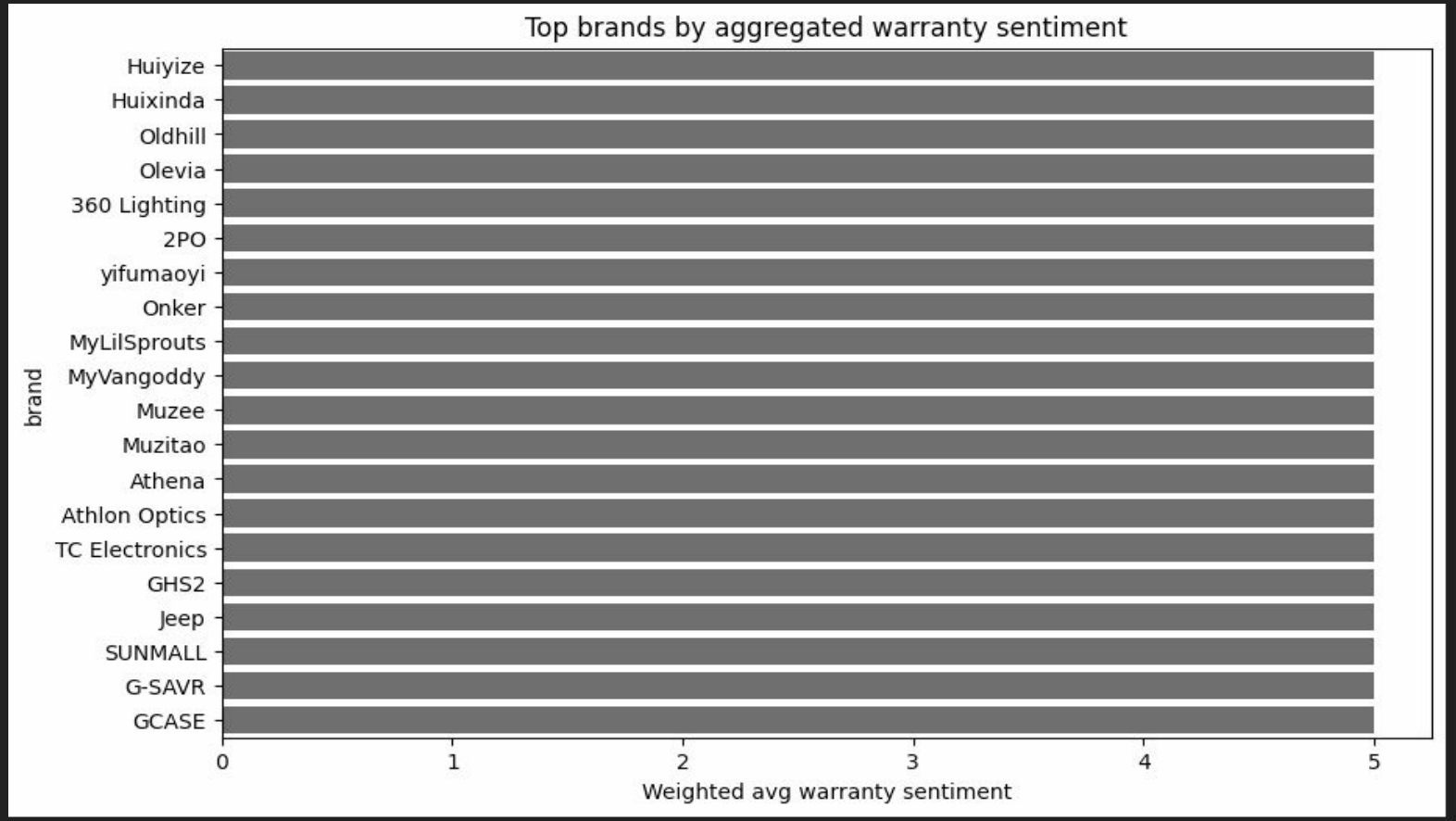
Top 10 Most Related Reviews to “Warranty”

Index	fullText	warranty_semantic_score
71118	great customer service should not end once the...	0.689735
108535	had a repair done and it was a fast and easy p...	0.676736
381327	great customer service . great customer servic...	0.658046
131615	bad return policy. product itself okay, but to...	0.654404
11103	a dead on arrival b got this confused with the...	0.654165
371182	i am really really glad i buy these warranties...	0.651214
63300	no returns allowed!!! . one star because the ...	0.640888
122238	pray you never need warranty service! . router...	0.639816
25780	arrived in good condition and looks attractive...	0.626262
15293	no return policy item does not fit the device...	0.622462

Top 10 Least Related Reviews to “Warranty”

Index	fullText	warranty_semantic_score
128967	will not work with all switches . i found that...	0.0
128966	not built to last . longevity was 4 months. i ...	0.0
128965	not great. reverts to static regularly. . this...	0.0
128964	not a fan . the rca lines are not built to be ...	0.0
128963	battery life need improvement . i have only ha...	0.0
128962	flat cords have built in holder but dont last ...	0.0
128961	hasnt been dependable . im really disappointed...	0.0
128960	terrible sound quality for music . these have ...	0.0
128959	seems to stop working after a short period . t...	0.0
128970	nope, just...no. . nope, just...no. where to s...	0.0

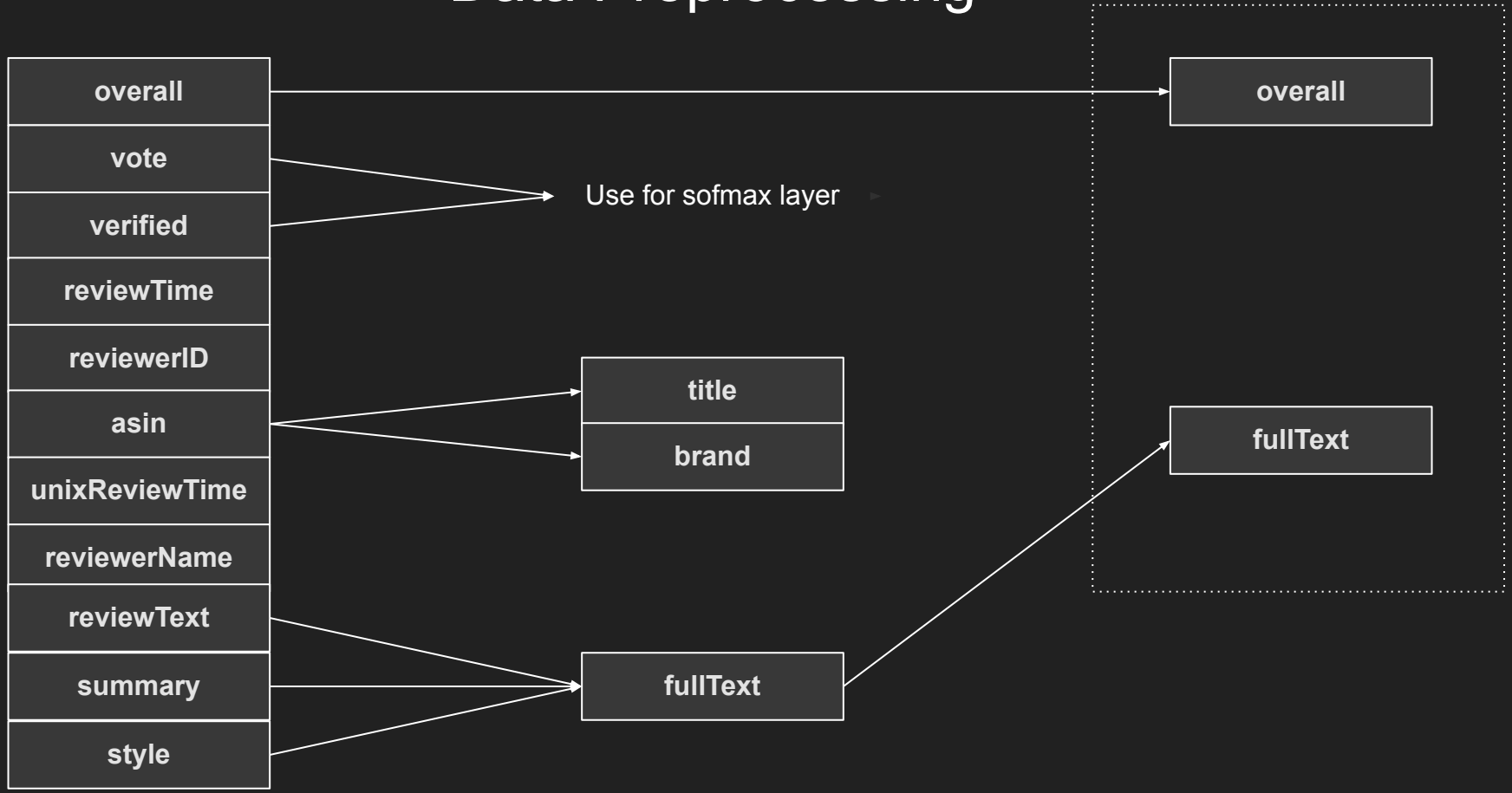
Warranty Satisfaction Ranking: The Top 20 Brands



Amazon reviews sentiment analysis

Predict sentiment of user reviews
(1–5 rating)

Data Preprocessing



Model Choice

DeBERTa v3 base

why?

Content + position

High benchmarks

Large data

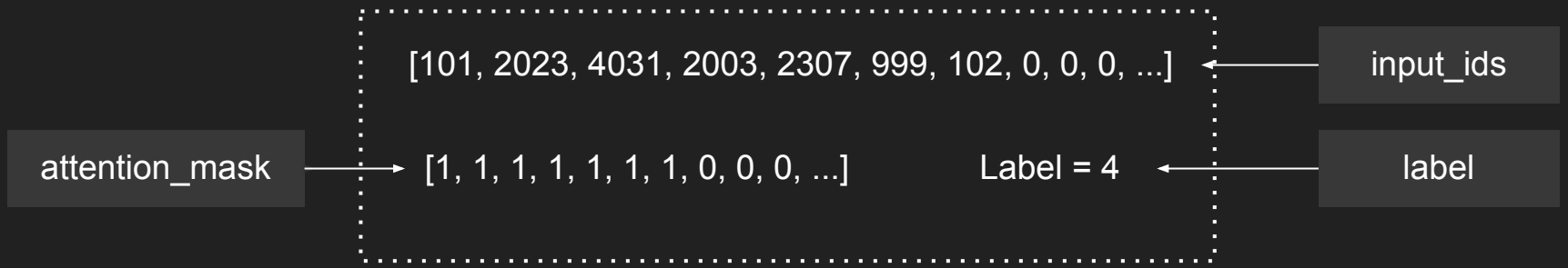
Tokenize

"This product is great!"

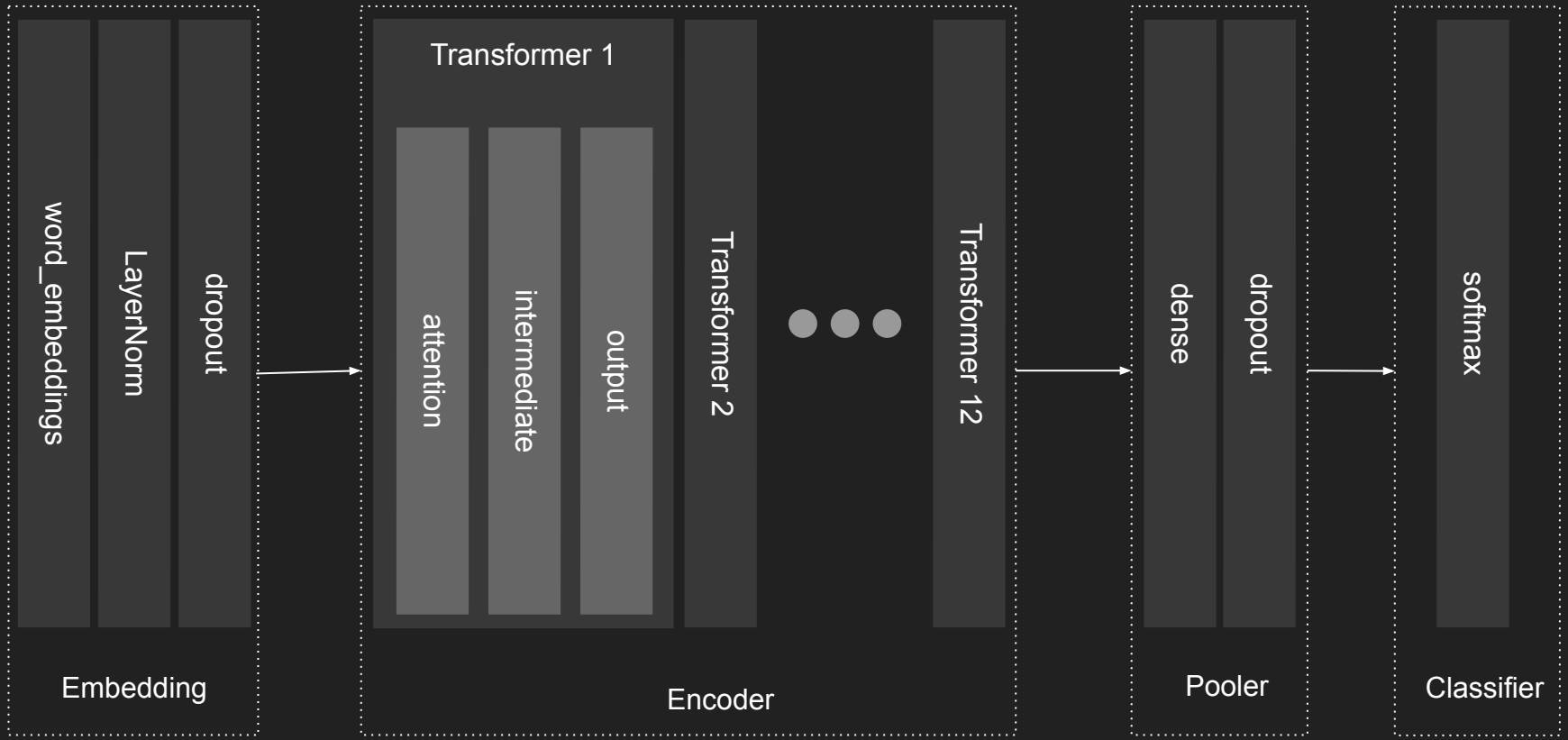
“this product is great!"

[“this” , ”product” , ”is” , ”great" , “!” , [SEP] , [CLS]]

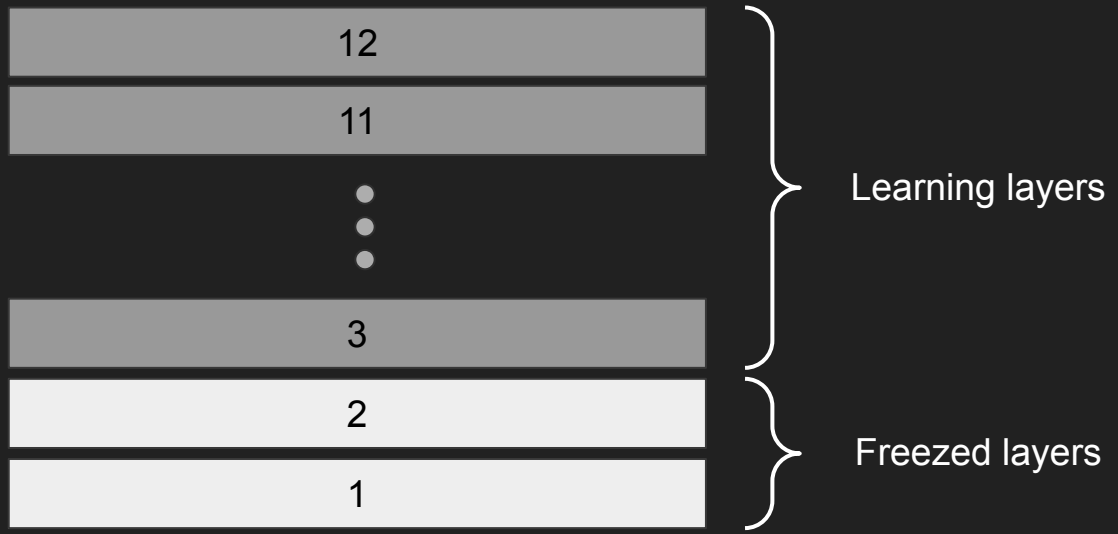
[101, 2023, 4031, 2003, 2307, 999, 102]



DeBERTa v3 base

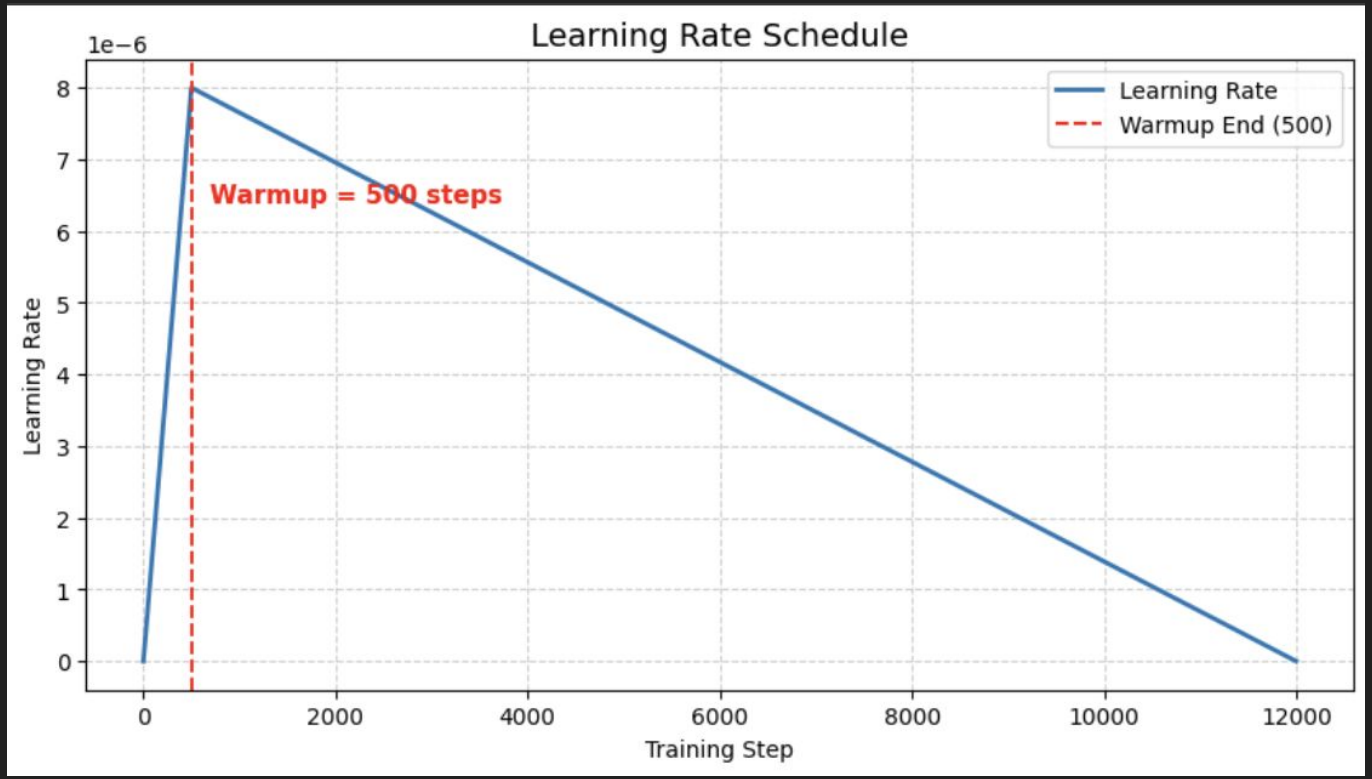


Freezing 2 layers

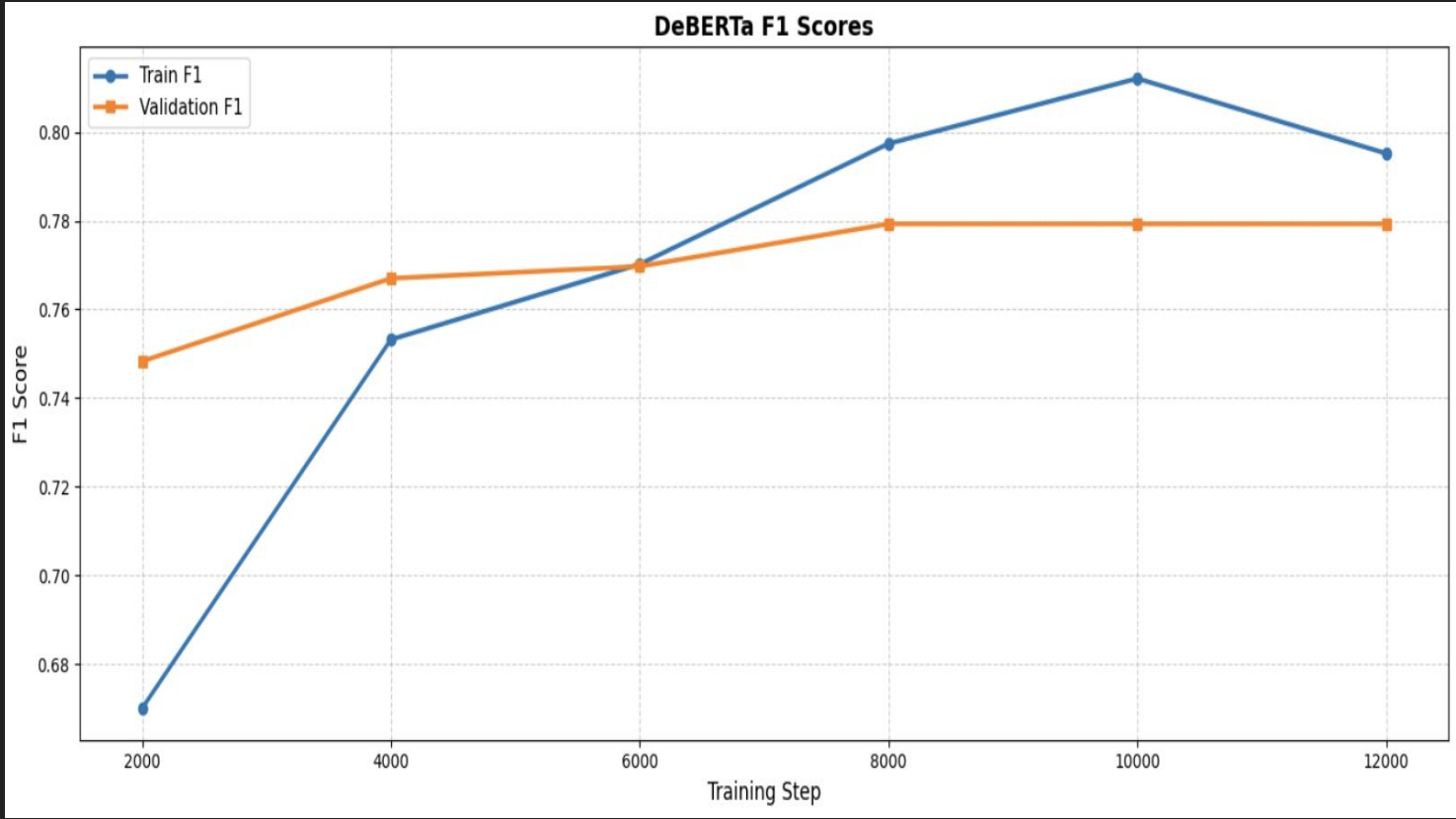


Scheduler

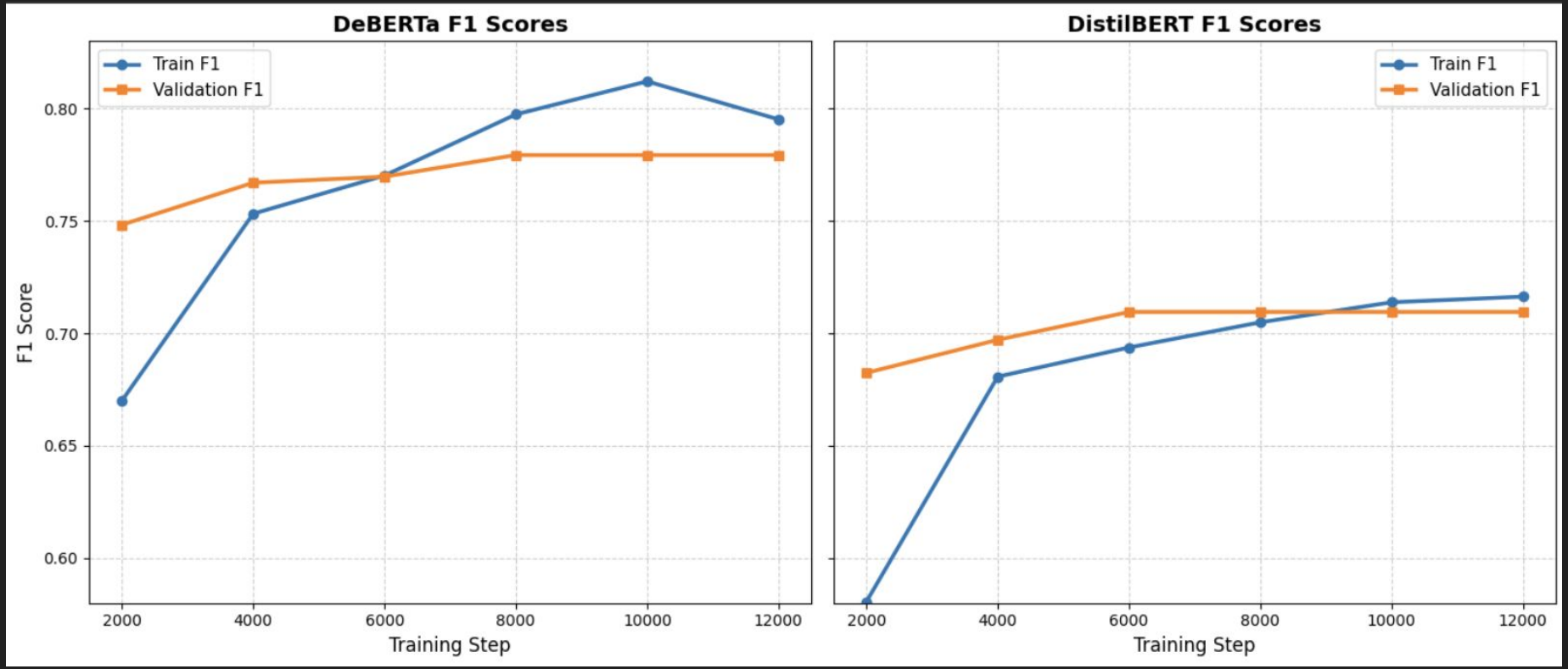
LEARNING_RATE = 8e-6 WARMUP_STEPS = 500



Result


DeBERTa F1 Scores
Train F1
Validation F1
F1 Score
0.80
0.78
0.76
0.74
0.72
0.70
0.68
2000
4000
6000
8000
10000
12000
Training Step

Result


DeBERTa F1 Scores
Train F1
Validation F1
F1 Score
0.80
0.75
0.70
0.65
0.60
2000
4000
6000
8000
10000
12000
Training Step
DistilBERT F1 Scores
Train F1
Validation F1
2000
4000
6000
8000
10000
12000
Training Step

Thanks