1.

" Clothing\_Shoes\_and\_Jewelry.csv "

Mismatched Count: 31119

Mismatched Percentage: 6.22%

Electronics.csv

Mismatched Count: 35786

Mismatched Percentage: 7.16%

Health\_and\_Personal\_Care.csv

Mismatched Count: 40943

Mismatched Percentage: 8.29%

3.

Clothing\_Shoes\_and\_Jewelry.csv

🧠 Top Keywords per Topic:

Topic 1: great, nice, good, fit, like, size, quality, cute, small, perfect

Topic 2: br, like, size, nice, fit, good, wear, just, comfortable, great

Topic 3: love, great, perfect, fit, comfortable, cute, soft, color, fits, shirt

Topic 4: stars, great, love, nice, comfortable, good, fit, quality, loves, cute

Topic 5: comfortable, shoes, great, fit, wear, shoe, good, soft, feet, nice

A graph with blue squares

AI-generated content may be incorrect.

📉 Average Rating Difference per Topic:

Topic Avg Rating Diff Review Count

0 0 -0.349472 310228

1 1 -0.575350 41971

2 2 -0.295603 43714

3 3 -0.157247 29972

4 4 -0.419891 59425

📝 Text Length Distribution per Topic (%):

length\_category Very Short Short Medium Long Very Long

topic

0 39.375288 32.647771 19.959118 7.152783 0.865040

1 6.111363 14.643444 34.442830 29.634748 15.167616

2 54.422329 34.587051 9.064287 1.814230 0.112102

3 98.628720 1.111037 0.226878 0.033364 0.000000

4 38.713586 35.256896 18.649972 6.538093 0.841453

✅ Verified Purchase Distribution per Topic (%):

Not Verified Verified

topic

0 19.110138 80.889862

1 58.190179 41.809821

2 11.490598 88.509402

3 3.653410 96.346590

4 21.337821 78.662179

A graph with blue rectangles

AI-generated content may be incorrect.

Thanks for sharing your results! Here's a **complete analysis** based on your TF-IDF + KMeans clustering output and plots:

**🧠 1. Topic Overview (Top Keywords)**

Your 5 clusters (topics) are generally **positive**, with keywords like:

* **"great", "good", "comfortable", "fit", "love", "perfect"**

This suggests the model grouped sentiment-heavy reviews, **but all topics skew positive**, which is common with real-world product reviews (positivity bias).

**📊 2. Topic Distribution**

| **Topic** | **Review Count** | **% of Total** |
| --- | --- | --- |
| 0 | 310,228 | ~62% |
| 4 | 59,425 | ~12% |
| 1 | 41,971 | ~8% |
| 2 | 43,714 | ~9% |
| 3 | 29,972 | ~6% |

🔍 **Topic 0 dominates** the dataset. This could be a catch-all for general short and vague praise like "nice", "fit", "good".

**📉 3. Rating Mismatch (Predicted – Actual)**

| **Topic** | **Avg Rating Diff** | **Comment** |
| --- | --- | --- |
| 3 | **-0.16** | Least mismatch — best aligned with text. |
| 2 | -0.30 | Slight underestimation |
| 0 | -0.35 | Text may sound better than rating |
| 4 | -0.42 | Moderate mismatch |
| 1 | **-0.58** | ⚠️ **Largest mismatch** — text is positive, but actual ratings are lower |

🧠 Interpretation:

* **All topics show underestimation**, meaning predicted stars are consistently lower than user-given ratings.
* Topic 1 could include **sarcastic**, **low-rated** reviews with positive words — worth exploring with examples.

**✏️ 4. Review Length Distribution**

| **Topic** | **% Very Short** | **% Long or Very Long** | **Interpretation** |
| --- | --- | --- | --- |
| 3 | 98.6% | ~0% | Shortest — very brief blurbs |
| 1 | 6.1% | 44.8% | 💡 Most detailed and expressive |
| 2 | 54.4% | 1.9% | Mostly short |
| 0 | 39.4% | 8% | Mixed |
| 4 | 38.7% | 7.4% | Mixed |

🧠 Topic 1 has **the longest reviews** — and the **largest rating mismatches**. This could mean:

* Users give long, nuanced reviews but **give lower stars** than the tone suggests.
* The model struggles to fully interpret the sentiment of long reviews.

**✅ 5. Verified Purchase Distribution**

| **Topic** | **% Verified** | **Interpretation** |
| --- | --- | --- |
| 3 | 96% | Very authentic |
| 2 | 88% | Reliable |
| 0 | 81% | Strong base |
| 4 | 78% | Good trust |
| 1 | **42%** | ⚠️ Least verified — may contain **more spammy or fake reviews** |

📌 Topic 1 again stands out: long reviews, high mismatch, and low verification — possibly **off-topic**, **biased**, or **deceptive** reviews.

**🔎 Overall Insights**

| **Topic** | **Traits** |
| --- | --- |
| **0** | Most common, short-to-medium, moderately mismatched |
| **1** | Longest reviews, **largest rating gap**, **lowest verified %** — investigate! |
| **2** | Short-to-medium reviews, decent alignment |
| **3** | Shortest reviews, best alignment |
| **4** | Mixed length, moderately mismatched |

**💡 Suggestions:** Longest reviews, **largest rating gap**, **lowest verified %** — investigate!

* Sample reviews from **Topic 1** to manually verify the mismatches.
* Consider **training a new sentiment model** that better handles:
  + **Sarcasm**
  + **Long reviews**
  + **Mixed tone (e.g., praise + criticism)**

Want help extracting review samples from Topic 1 or visualizing any of this in a report?

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force\_remount=True).

351A2 IMG\_1415.MOV

Clothing\_Shoes\_and\_Jewelry.csv 'Untitled document (1).gdoc'

'Colab Notebooks' 'Untitled document.gdoc'

'Copy of Electronics.csv' 杨立行出发

'Copy of Health\_and\_Personal\_Care.csv' 杨立行出发.zip

🧠 Top Keywords per Topic:

Topic 1: great, nice, good, fit, like, size, quality, cute, small, perfect

Topic 2: br, like, size, nice, fit, good, wear, just, comfortable, great

Topic 3: love, great, perfect, fit, comfortable, cute, soft, color, fits, shirt

Topic 4: stars, great, love, nice, comfortable, good, fit, quality, loves, cute

Topic 5: comfortable, shoes, great, fit, wear, shoe, good, soft, feet, nice

A graph with blue squares

AI-generated content may be incorrect.

📉 Average Rating Difference per Topic:

Topic Avg Rating Diff Review Count

0 0 -0.349472 310228

1 1 -0.575350 41971

2 2 -0.295603 43714

3 3 -0.157247 29972

4 4 -0.419891 59425

📝 Text Length Distribution per Topic (%):

length\_category Very Short Short Medium Long Very Long

topic

0 39.375288 32.647771 19.959118 7.152783 0.865040

1 6.111363 14.643444 34.442830 29.634748 15.167616

2 54.422329 34.587051 9.064287 1.814230 0.112102

3 98.628720 1.111037 0.226878 0.033364 0.000000

4 38.713586 35.256896 18.649972 6.538093 0.841453

✅ Verified Purchase Distribution per Topic (%):

Not Verified Verified

topic

0 19.110138 80.889862

1 58.190179 41.809821

2 11.490598 88.509402

3 3.653410 96.346590

4 21.337821 78.662179

A graph with blue rectangles

AI-generated content may be incorrect.

review\_text predict\_stars rating

0 I SHOULD'VE KNOWN This chain is not 18", it fa... 3 1

1 Size Up! Well made and good materials are used... 3 1

2 Not very flattering; fits small I was disappoi... 3 1

3 Updated review - Hunk of Junk Updated review ... 3 1

4 Pearl is plastic with a nice luster. Cheesy. ... 3 1

5 Glue When the shoes came they smelled like sho... 4 2

6 I got a medium I’m 5’6” 140lbs and moderately ... 3 1

7 Can Never Find A 48B That Will Fit I ordered a... 3 1

8 Soxart Athletic Socks These Soxart Athletic So... 4 2

9 Not My Style!!! I Do Not Like The Item Design ... 3 1

10 Cute - Adorable, but didn't fit. Bought my reg... 3 1

11 Classic style My boyfriend is 6' - 4" and I ty... 3 1

12 Weirdest cut I've ever seen! Awkward, ill-fitt... 3 1

13 Polar bear feet! Don't buy if you have an inst... 3 1

14 Oddly sized Ok, so these are sized strangely. ... 4 2

<ipython-input-18-efff78b2bb22>:118: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(verified['text\_length'], label='Verified', shade=True)

<ipython-input-18-efff78b2bb22>:119: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(unverified['text\_length'], label='Unverified', shade=True)

/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151: UserWarning: Glyph 128221 (\N{MEMO}) missing from font(s) DejaVu Sans.

fig.canvas.print\_figure(bytes\_io, \*\*kw)

A graph with a number of words

AI-generated content may be incorrect.

<ipython-input-18-efff78b2bb22>:127: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(verified['rating\_diff'], label='Verified', shade=True)

<ipython-input-18-efff78b2bb22>:128: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(unverified['rating\_diff'], label='Unverified', shade=True)

/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151: UserWarning: Glyph 128201 (\N{CHART WITH DOWNWARDS TREND}) missing from font(s) DejaVu Sans.

fig.canvas.print\_figure(bytes\_io, \*\*kw)

A graph of a number of stars

AI-generated content may be incorrect.

A diagram of a bar graph

AI-generated content may be incorrect.

Group Misleading Rate (%) Review Count

0 Long + Unverified 7.972270 2308

1 Long + Verified 12.026359 1214

2 Short + Unverified 5.142857 175

3 Short + Verified 8.665511 577

A graph of misleading rate

AI-generated content may be incorrect.

<ipython-input-18-efff78b2bb22>:182: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(data=topic\_1\_reviews[topic\_1\_reviews['length\_category'] == 'Long'],

<ipython-input-18-efff78b2bb22>:184: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(data=topic\_1\_reviews[topic\_1\_reviews['length\_category'] == 'Short'],

/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151: UserWarning: Glyph 128221 (\N{MEMO}) missing from font(s) DejaVu Sans.

fig.canvas.print\_figure(bytes\_io, \*\*kw)

A graph of a number of stars

AI-generated content may be incorrect.

<ipython-input-18-efff78b2bb22>:212: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(data=subset, x='rating\_diff', label=group, shade=True, color=color)

<ipython-input-18-efff78b2bb22>:212: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(data=subset, x='rating\_diff', label=group, shade=True, color=color)

<ipython-input-18-efff78b2bb22>:212: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(data=subset, x='rating\_diff', label=group, shade=True, color=color)

<ipython-input-18-efff78b2bb22>:212: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(data=subset, x='rating\_diff', label=group, shade=True, color=color)

/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151: UserWarning: Glyph 128201 (\N{CHART WITH DOWNWARDS TREND}) missing from font(s) DejaVu Sans.

fig.canvas.print\_figure(bytes\_io, \*\*kw)

A graph of a number of stars

AI-generated content may be incorrect.

<ipython-input-18-efff78b2bb22>:239: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(data=subset, x='rating\_diff', label=group, shade=True, color=color)

<ipython-input-18-efff78b2bb22>:239: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(data=subset, x='rating\_diff', label=group, shade=True, color=color)

<ipython-input-18-efff78b2bb22>:239: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(data=subset, x='rating\_diff', label=group, shade=True, color=color)

<ipython-input-18-efff78b2bb22>:239: FutureWarning:

`shade` is now deprecated in favor of `fill`; setting `fill=True`.

This will become an error in seaborn v0.14.0; please update your code.

sns.kdeplot(data=subset, x='rating\_diff', label=group, shade=True, color=color)

/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151: UserWarning: Glyph 128201 (\N{CHART WITH DOWNWARDS TREND}) missing from font(s) DejaVu Sans.

fig.canvas.print\_figure(bytes\_io, \*\*kw)

A graph of a number of stars

AI-generated content may be incorrect.

1 to 5 of 5 entriesFilter

| **index** | **review\_text** | **predict\_stars** | **rating** |
| --- | --- | --- | --- |
| **0** | Pretty This is a cute dress. I liked it. I thought it was flattering. I thought the fit was pretty good. I thought that it was unique and looked good.<br /><br />The dress was relatively comfortable. The fit was okay. It wasn't too hot, which for me is important since we live in Florida.<br /><br />It wasn't too hard to clean. I thought it seemed trendy. It felt durable and the fabric was relative... | 3 | 5 |
| **1** | 60% cotton / 40% polyester. Soft and comfortable The top of this set is slightly softer than the bottom. But neither are “rough” in any way. The top is made of 60% cotton and 40% polyester. I have the size Small, and it measures 23” from the top of the shoulder to the bottom. It measures 19” at the midsection and at the bottom.<br /><br />The bottom is also made of 60% cotton and 40% polyeste... | 3 | 5 |
| **2** | I agree with the assessment that this is for C cups maximum I have a somewhat larger bust and I was really hoping that this cute little bra would work for comfy days around the house or as a pajama bra. Unfortunately, those spaghetti straps are just a little too thin to provide any support at all to a larger bust.<br /><br />However, the fabric and the fit overall is really lovely. It is soft, has... | 3 | 5 |
| **3** | Runs a Bit Small This top is really cute. The double ruffle is feminine and flattering. The fabric is lightweight and has a crinkle texture. The fit runs small. I couldn't pull this shirt all the way down. It was too snug over my hips. You may want to size up if you want to wear it untucked. For fit reference, I'm 5'6", 153lbs and a 34G bra size. I'm wearing the medium.<br /><br />Thanks so much f... | 3 | 5 |
| **4** | This Rain jacket has everything to keep you dry.. Outdoor Ventures thought of Everything to keep you dry and comfortable with this Jacket.<br />Plus it’s good looking.<br />Pockets on both sides, and chest with zippers. Inside small ziper pocket.<br />Velcro pulls on each wrist to keep you dry.<br />On large hood in the back there is an elastic pull to make the hood smaller and tighter, And Two mo... |  |  |

Electronics.csv

A graph of a number of stars

AI-generated content may be incorrect.

Health\_and\_Personal\_Care.csv

A graph of a number of stars

AI-generated content may be incorrect.