

۱ فاز اول: استخراج ویژگی‌ها

۱.۱ چالش‌های داده‌های خام

- وجود نویزهایی مانند نشانی‌های وب، ایمیل، علائم نگارشی و تکرار حروف
- شامل بودن کلمات پرتکرار (stopwords) که اطلاعات مفیدی منتقل نمی‌کنند
- وجود کلمات با فرمت‌های گرامری متفاوت (مانند run، running، ran) که باید به یک ریشه نگاشته شوند
- وجود حروف بزرگ/کوچک که می‌توانند مدل را دچار سردرگمی کنند (مثلاً Apple با apple متفاوت تفسیر شود)

۲.۱ راهکارها و منطق انتخاب‌شده

دلایل هر تکنیک استفاده شده:

- **تبدیل به حروف کوچک (lowercase):** برای یکسان‌سازی نوشتار کلمات و جلوگیری از تکرار غیر ضروری ویژگی‌ها.
- **حذف لینک‌ها و ایمیل‌ها:** این اطلاعات معمولاً حاوی معنای خاصی برای دسته‌بندی محصول نیستند و حذف آن‌ها به ساده‌سازی متن کمک می‌کند.
- **حذف اعداد:** اعداد در این نوع مسئله‌ها معمولاً کاربرد خاصی ندارند (مثلاً شماره مدل یا سایز)، بنابراین حذف شدند مگر اینکه در تحلیل آینده به صورت جدا استفاده شوند.
- **حذف علائم نگارشی و HTML:** این کار برای کاهش نویز و تمرکز بر محتوای متنی اصلی انجام شد.
- **کاهش کشیدگی کلمات:** کلماتی مثل soooo به so تبدیل شدند تا مدل دچار تکرار بی‌مورد نشود.
- **حذف کلمات توقف (Stopwords):** کلماتی مثل in، is، the، اطلاعات خاصی منتقل نمی‌کنند و حضور آن‌ها باعث می‌شود مدل یادگیری سخت‌تری داشته باشد.

- **Lemmatization:** برای تبدیل کلمات به شکل پایه‌شان (مثل running به run) از Lemma-tizer استفاده شد. این کار به مدل کمک می‌کند مفهوم اصلی کلمه را فارغ از فرم گرامری‌اش یاد بگیرد.

- **توکن‌سازی:** برای پردازش بهتر، متن به کلمات جداگانه شکسته شد تا بتوان روی هر کلمه عملیات انجام داد.

در پایان هر نمونه متنی به یک نسخه تمیز، یکنواخت و کاهش‌یافته از نظر اطلاعات زائد تبدیل شد. این داده‌ها اکنون آماده‌اند تا در مراحل بعدی برای برداری‌سازی (با TF-IDF و Word2Vec) استفاده قرار گیرند.

	text	clean_text
35848	Kandy Men's Regular Fit Blazer Blue This produ...	kandy men regular fit blazer blue product made...
13005	HealthSense Chef-Mate KS 50 Digital Kitchen Sc...	healthsense chefmate digital kitchen scale grey
22719	Concept of Physics (2018-2019) Session (Set of...	concept physic session set volume
18453	Lista Stainless Steel Multi Functional Hammer ...	lista stainless steel multi functional hammer ...
20867	Gardening in Urban India update	gardening urban india update

۲ فاز دوم: بردارسازی متون با استفاده از TF-IDF

در این بخش، الگوریتم TF-IDF بدون استفاده از هیچ کتابخانه‌ای پیاده‌سازی شد.

۱.۲ چالش‌ها و راهکارها

- **محاسبه DF:** یکی از چالش‌ها، محاسبه دقیق Frequency Document برای هر واژه بود، به طوری که شمارش تکرار یک واژه فقط یک‌بار در هر سند لحاظ شود. برای حل این مسئله، از set استفاده شد تا کلمات تکراری در هر سند فقط یک‌بار شمرده شوند.

- **جلوگیری از تقسیم بر صفر:** در محاسبه IDF، برای جلوگیری از تقسیم بر صفر در فرمول $\log(N/df)$ ، یک واحد به df اضافه شد، یعنی از $\log(N/(1 + df))$ استفاده شد.

- **ابعاد ماتریس:** با توجه به تعداد واژگان یکتا (واژگان نهایی)، لازم بود که ابتدا دیکشنری word2idx تعریف شود تا اندیس هر واژه در ماتریس TF-IDF مشخص شود. سپس یک ماتریس $N \times V$ (تعداد اسناد \times تعداد واژگان) ساخته شد.

۲.۲ نتایج

نتیجه داکيومنت اول نشان داد که واژگانی مانند kandy، velvet و buttoned دارای مقادیر TF-IDF بالاتری نسبت به کلمات عمومی‌تر مانند product یا made هستند. این نشان می‌دهد که TF-IDF به درستی کلمات متمایزکننده را نسبت به کلمات پرتکرار تشخیص داده است.

```
kandy men regular fit blazer blue product made velvet finished attractive blue color feature plain solid pattern long sle
kandy → tf-idf: 7.824046010856292
men → tf-idf: 2.8134107167600364
regular → tf-idf: 3.519980917652122
fit → tf-idf: 2.501036031717884
blazer → tf-idf: 5.626821433520073
blue → tf-idf: 2.9603651297166995
product → tf-idf: 1.7070507413011007
made → tf-idf: 1.789761466565382
velvet → tf-idf: 5.8781358618009785
finished → tf-idf: 4.06284589516273
attractive → tf-idf: 3.519980917652122
blue → tf-idf: 2.9603651297166995
color → tf-idf: 1.9589953886039688
feature → tf-idf: 2.0826466726287842
plain → tf-idf: 4.406319327242926
solid → tf-idf: 3.649658740960655
pattern → tf-idf: 3.4357888264317746
long → tf-idf: 2.4557362724882204
sleeve → tf-idf: 3.789805372703897
buttoned → tf-idf: 6.907755278982137
closure → tf-idf: 5.051457288616511
targeted → tf-idf: 5.8781358618009785
towards → tf-idf: 5.149897361429764
men → tf-idf: 2.8134107167600364
furthermore → tf-idf: 5.115995809754082
recommended → tf-idf: 4.173387769562553
kept → tf-idf: 5.221356325411908
away → tf-idf: 3.9220733412816475
extreme → tf-idf: 5.051457288616511
heat → tf-idf: 3.533586569707901
fire → tf-idf: 5.339139361068292
corrosive → tf-idf: 5.8781358618009785
liquid → tf-idf: 4.474141923581687
avoid → tf-idf: 4.268697949366879
form → tf-idf: 3.763603000309873
damage → tf-idf: 4.06284589516273
```

۳ فاز دوم: بردارسازی متون با استفاده از Word2Vec

از مدل Word2Vec کتابخانه gensim استفاده شد.

۱.۳ چالش‌ها و تصمیمات

• توکن‌سازی دقیق: برای آموزش Word2Vec نیاز به توکن‌سازی دقیق داشتیم. در ابتدا از split() استفاده شد، اما برای نتایج بهتر از word_tokenize از کتابخانه nltk استفاده شد.

• تنظیمات مدل: برای آموزش مدل، پارامترهایی مانند window=5، vector_size=100 و min_count=2 انتخاب شد. این تنظیمات با توجه به اندازه داده‌ها و برای جلوگیری از نویز ناشی از کلمات بسیار کم‌تکرار انجام شد.

• **کلمات نادیده گرفته شده:** برخی کلمات به دلیل min_count وارد مدل نمی‌شوند، بنابراین در تحلیل نهایی ممکن است برخی واژه‌ها بردار نداشته باشند.

۲.۳ نتایج

با بررسی بردار کلمه painting مشاهده شد که شباهت زیادی با واژگان مفهومی مانند art، drawing، و craft دارد. این نشان می‌دهد که مدل Word2Vec توانسته ارتباط معنایی بین کلمات را تا حد مناسبی بیاموزد. همچنین خروجی بردار عددی کلمه painting به خوبی ساختار توزیعی مدل را نشان می‌دهد.

```
Vector for 'painting':
[-0.5508629  0.36387342 -0.36982006 -1.056777  -0.22129601 -0.11388729
 -1.4613373  -0.73587257 -1.2004316  2.6038702  0.2659391 -0.6010603
 1.0231673  -0.92522097 -0.79301536 -1.2032902  1.4779582  2.5695066
 -0.15849325  1.3117727  -0.2737219  -0.61334205 -0.5723068  0.25922787
 0.81314796  0.60878175  -0.35792628  -0.6534402  0.7085712  -1.3418529
 -0.53413486  -0.42737135  0.68361074  -0.19693144  -0.2622345  -0.19879907
 -2.0610995  -0.6067788  -0.26163772  -0.28893083  0.2548341  -1.48044
 -2.6918893  0.86935455  -1.7898625  0.23239349  -1.2008233  -1.0039829
 1.537418  -1.3866178  -1.1436353  -0.8453917  -1.0944076  0.17852592
 0.9740102  0.05182031  -0.6212762  -1.4812379  -1.5863237  -0.35590577
 0.42921606  -1.2403116  0.48564795  -1.1095368  0.5501702  -1.0542358
 -0.35811844  1.2013218  -1.0504943  1.2214376  -0.88091445  0.7849959
 0.2985102  -0.42124787  0.01212086  -0.2500986  0.52507186  -0.41478604
 0.80750847  1.1138465  -0.0455182  -1.648858  -1.3382894  1.0623164
 -0.88975775  -0.12223998  0.5895954  2.7580974  -0.5656033  -0.42690265
 -0.60894865  3.018483  0.7489087  0.8160427  0.4866975  2.2931328
 -1.6081651  0.40424594  -0.9389486  0.14996298]
```

```
Similarity between 'painting' and 'art': 0.6907795071601868
Most similar words to 'painting':
art: 0.6907795071601868
watercolour: 0.6684693098068237
craft: 0.6585366129875183
drawing: 0.6484317779541016
stencil: 0.6409387588500977
```

۴ فاز سوم و چهارم

در فاز سوم بردار های به دست آمده از فاز قبل را طبقه بندی کردیم، سپس عملکرد دو روش رو بررسی کردیم.

(ابتدا تلاش کردیم که نتایج فاز قبلی را ذخیره کنیم که دوباره کل کد را اجرا نکنیم ولی خطای ram system را می‌گرفتیم که با کاهش دیتا ست به ۵۰۰۰ تا این خطا رفع شد) برای طبقه بندی از سه مدل linear svc , logistic regression , random forest استفاده کردیم. شبکه عصبی را هم امتحان کردیم ولی زمان زیادی لازم داشت و نتایج بقیه مطلوب بود به همین دلیل

آن را حذف کردیم. و نتایج نشان میداد که Logistic Regression بهتر و پایدارتر در هر دو عمل کرده پس از نتایج آن استفاده کردیم و ماتریس سردرگمی را هم برای این مدل رسم کردیم.

۱.۴ مقایسه کلی مدل‌ها

Accuracy Word2Vec	Accuracy TF-IDF	مدل
0.906	0.94	Logistic Regression
0.909	0.944	Linear SVC
0.916	0.921	Random Forest

مقایسه دقت مدل‌ها بر اساس نمایش ویژگی‌ها

با توجه به نتایج مدل Linear SVC بالاترین دقت را بر روی TF-IDF کسب کرده است. با این حال، مدل Logistic Regression در هر دو نمایش عملکردی نسبتاً بالا و پایدار داشته و تفاوت کمی بین دو نمایش دارد، در حالی که SVC Linear و Forest Random افت بیشتری در دقت روی Word2Vec دارند. این ثبات دلیلی بر انتخاب نهایی Logistic Regression به عنوان مدل برتر است.

Random Forest

Accuracy: 0.921

	precision	recall	f1-score	support
Books	0.91	0.93	0.92	237
Clothing & Accessories	0.98	0.93	0.95	189
Electronics	0.93	0.87	0.90	202
Household	0.90	0.94	0.92	372
accuracy			0.92	1000
macro avg	0.93	0.92	0.92	1000
weighted avg	0.92	0.92	0.92	1000

Linear SVC

Accuracy: 0.944


	precision	recall	f1-score	support
Books	0.95	0.92	0.94	237
Clothing & Accessories	0.97	0.96	0.97	189
Electronics	0.90	0.93	0.92	202
Household	0.95	0.95	0.95	372
accuracy			0.94	1000
macro avg	0.94	0.94	0.94	1000
weighted avg	0.94	0.94	0.94	1000

Logistic Regression


Accuracy: 0.94

	precision	recall	f1-score	support
Books	0.97	0.91	0.94	237
Clothing & Accessories	0.98	0.94	0.96	189
Electronics	0.93	0.91	0.92	202
Household	0.91	0.97	0.94	372
accuracy			0.94	1000
macro avg	0.95	0.93	0.94	1000
weighted avg	0.94	0.94	0.94	1000


نتایج مدل های طبقه بندی برای tfidf

 Random Forest
Accuracy: 0.916

	precision	recall	f1-score	support
Books	0.96	0.91	0.93	237
Clothing & Accessories	0.94	0.90	0.92	189
Electronics	0.88	0.92	0.90	202
Household	0.89	0.93	0.91	372
accuracy			0.92	1000
macro avg	0.92	0.91	0.92	1000
weighted avg	0.92	0.92	0.92	1000

 Linear SVC
Accuracy: 0.909

	precision	recall	f1-score	support
Books	0.93	0.90	0.92	237
Clothing & Accessories	0.91	0.94	0.93	189
Electronics	0.87	0.88	0.88	202
Household	0.91	0.91	0.91	372
accuracy			0.91	1000
macro avg	0.91	0.91	0.91	1000
weighted avg	0.91	0.91	0.91	1000

 Logistic Regression
Accuracy: 0.906

	precision	recall	f1-score	support
Books	0.95	0.91	0.93	237
Clothing & Accessories	0.92	0.91	0.92	189
Electronics	0.88	0.87	0.88	202
Household	0.88	0.92	0.90	372
accuracy			0.91	1000
macro avg	0.91	0.90	0.91	1000
weighted avg	0.91	0.91	0.91	1000

نتایج مدل های طبقه بندی برای word2vec

۲.۴ تحلیل مدل و ماتریس سردرگمی

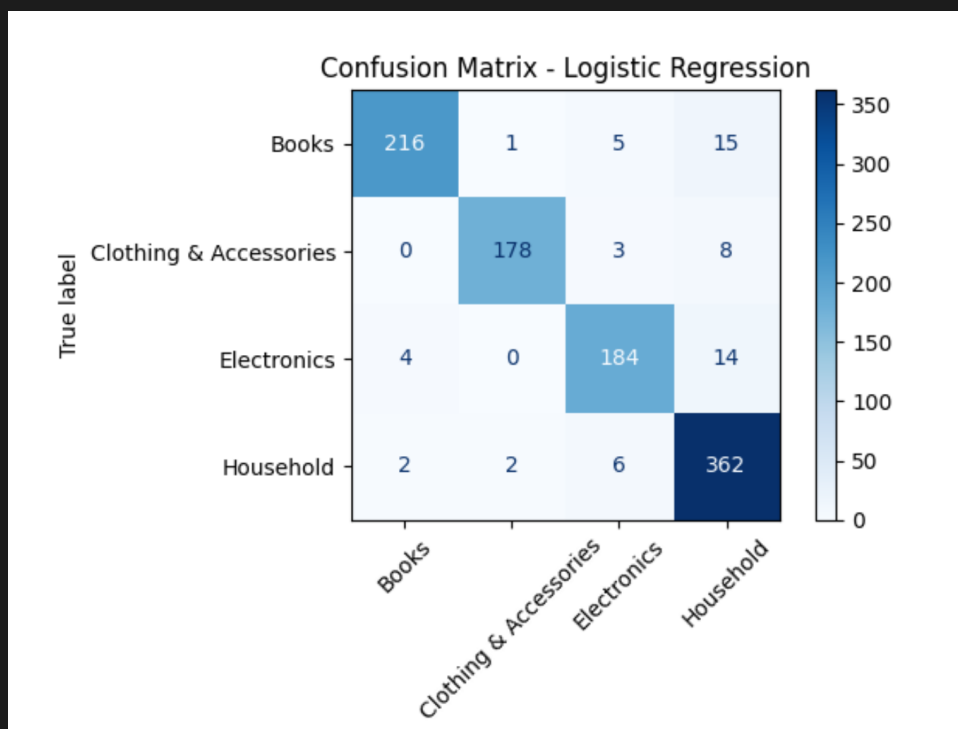
بررسی عملکرد مدل Regression Logistic

TF-IDF + Regression Logistic

Logistic Regression
Accuracy: 0.94

	precision	recall	f1-score	support
Books	0.97	0.91	0.94	237
Clothing & Accessories	0.98	0.94	0.96	189
Electronics	0.93	0.91	0.92	202
Household	0.91	0.97	0.94	372
accuracy			0.94	1000
macro avg	0.95	0.93	0.94	1000
weighted avg	0.94	0.94	0.94	1000

نتایج logistic regression برای tfidf



ماتریس سردرگمی برای tfidf

تحلیل:

- کتاب‌ها: دقت بسیار بالا (۹۷.۰) نشان می‌دهد که مدل به‌ندرت محصولات غیرکتاب را به اشتباه در این دسته قرار داده. اما recall پایین‌تر (۹۱.۰) بیانگر آن است که برخی کتاب‌ها به اشتباه در دسته‌های دیگر قرار گرفته‌اند.

- پوشاک و لوازم جانبی: دقت و recall بالا نشان‌دهنده‌ی توانایی بالای مدل در تشخیص صحیح این دسته است.
- الکترونیک: عبا وجود اینکه دقت و recall کمی پایین‌تر از پوشاک هستند، مدل در تشخیص این دسته نیز موفق عمل کرده است.
- لوازم خانگی: recall بسیار بالا (۹۷.۰) نشان می‌دهد که مدل تقریباً همه‌ی نمونه‌های این دسته را شناسایی کرده، اما دقت پایین‌تر (۹۱.۰) نشان‌دهنده‌ی آن است که برخی محصولات دیگر به اشتباه در این دسته قرار گرفته‌اند.

در مجموع:

- مدل در دسته‌ی پوشاک و لوازم جانبی عملکرد بسیار خوبی دارد و تقریباً همه نمونه‌ها را درست تشخیص داده است.
- در دسته‌ی کتاب‌ها، دقت بالا است اما recall کمی پایین‌تر است؛ یعنی برخی کتاب‌ها به اشتباه در دسته‌های دیگر قرار گرفته‌اند.
- الکترونیک نیز عملکرد خوبی دارد، اما نسبت به سایر دسته‌ها کمی ضعیف‌تر است.
- لوازم خانگی بیشترین میزان recall را دارد؛ یعنی مدل تقریباً همه‌ی نمونه‌های واقعی این دسته را شناسایی کرده است.
- بیشترین اشتباه‌ها بین دسته‌های کتاب - لوازم خانگی و الکترونیک - لوازم خانگی رخ داده‌اند.
- این اشتباه‌ها احتمالاً به دلیل شباهت متنی (و یا توصیفی) بین محصولات باشد (مثلاً کتاب‌های آشپزی یا وسایل دیجیتال خانگی).
- دسته‌ی پوشاک کمترین میزان اشتباه را دارد و مدل در تشخیص آن بسیار دقیق عمل کرده است.

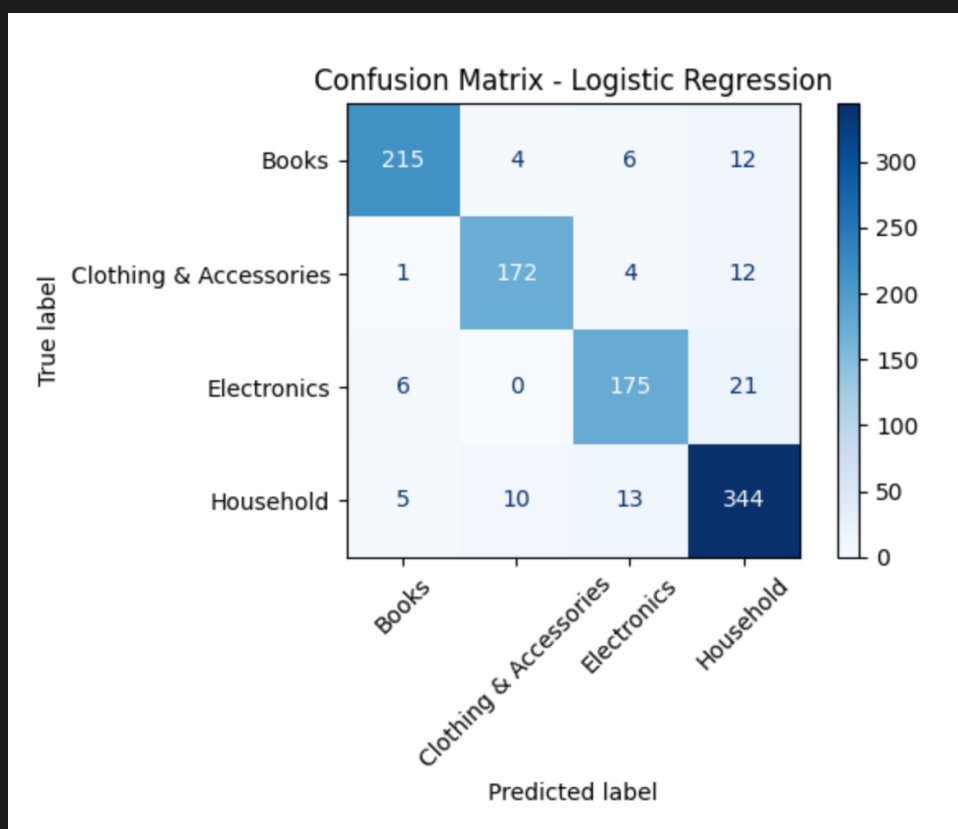
Word2Vec + Regression Logistic

Logistic Regression

Accuracy: 0.906

	precision	recall	f1-score	support
Books	0.95	0.91	0.93	237
Clothing & Accessories	0.92	0.91	0.92	189
Electronics	0.88	0.87	0.88	202
Household	0.88	0.92	0.90	372
accuracy			0.91	1000
macro avg	0.91	0.90	0.91	1000
weighted avg	0.91	0.91	0.91	1000

نتایج logistic regression برای tfidf



ماتریس سردرگمی برای tfidf

تحلیل:

• عملکرد ضعیف‌تر نسبت به TF-IDF در همه‌ی کلاس‌ها دارد.

- کتاب‌ها: دقت بالا (۹۵.۰) یعنی مدل به‌ندرت محصولات غیرکتاب را به اشتباه در این دسته قرار داده. اما recall پایین‌تر (۹۱.۰) نشان می‌دهد که برخی کتاب‌ها به اشتباه در دسته‌های دیگر افتاده‌اند.

- پوشاک: دقت و recall تقریباً برابرند، که نشان‌دهنده‌ی ثبات مدل در این دسته است.
- الکترونیک: ضعیف‌ترین عملکرد را دارد. هم دقت و هم recall پایین‌تر از سایر دسته‌ها هستند، که نشان می‌دهد مدل در تشخیص این دسته دچار سردرگمی است.
- لوازم خانگی: recall بالا (۹۲.۰) یعنی مدل بیشتر نمونه‌های این دسته را شناسایی کرده، اما دقت پایین‌تر (۸۸.۰) نشان می‌دهد که برخی محصولات دیگر به اشتباه در این دسته قرار گرفته‌اند.

در مجموع:

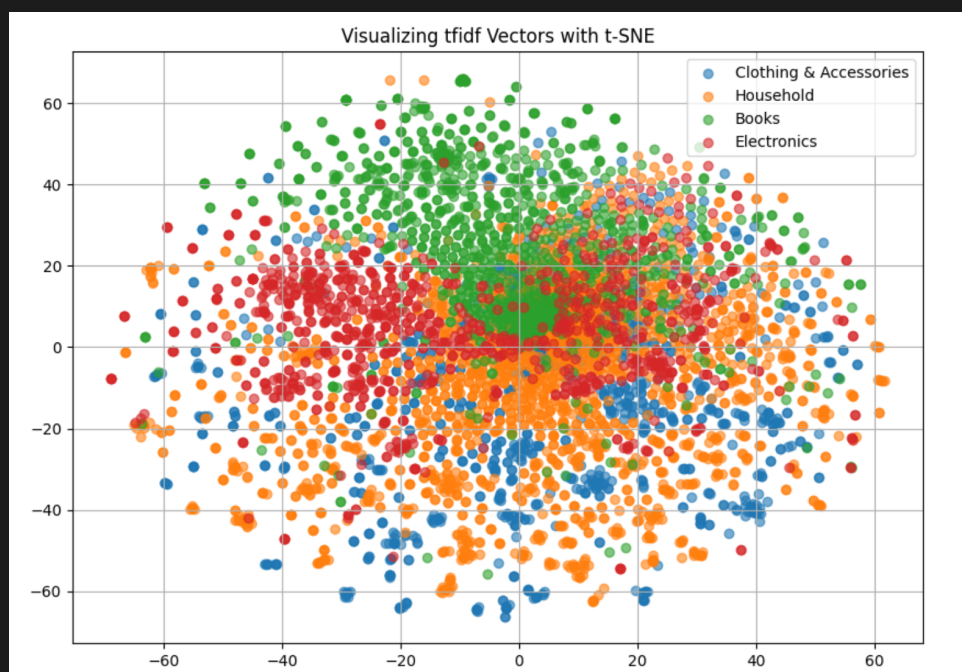
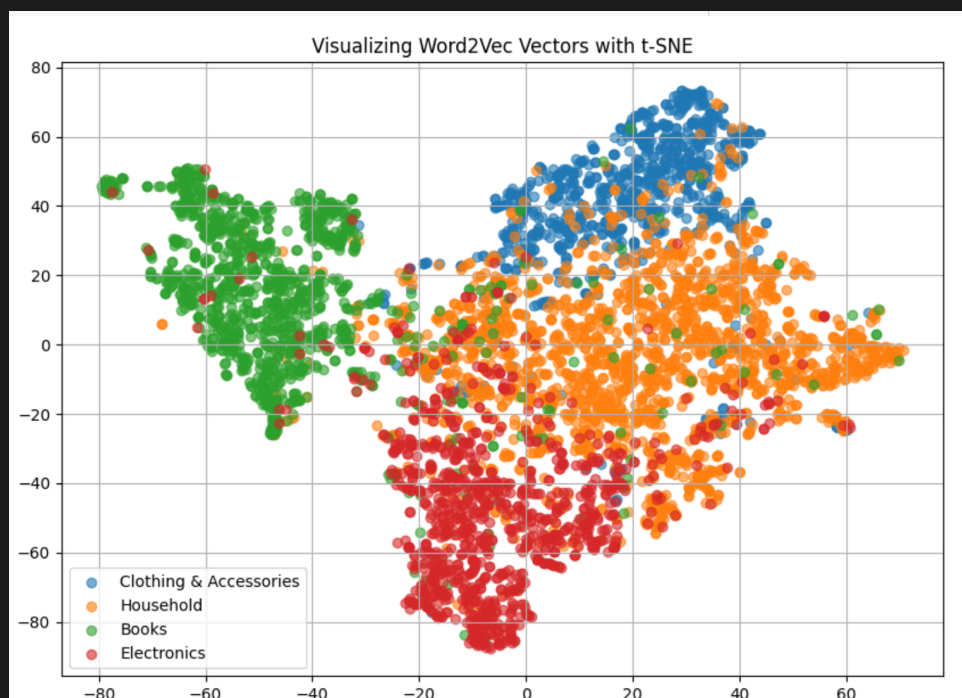
- مدل در دسته‌ی کتاب‌ها عملکرد بسیار خوبی دارد؛ دقت بالا و recall مناسب.
- پوشاک و لوازم جانبی نیز با دقت و recall نزدیک به هم، عملکردی پایدار دارد.
- الکترونیک ضعیف‌ترین عملکرد را دارد، به‌ویژه در recall که نشان‌دهنده‌ی اشتباه در تشخیص برخی نمونه‌هاست.
- لوازم خانگی با وجود دقت پایین‌تر، recall بالایی دارد؛ یعنی مدل بیشتر نمونه‌های واقعی این دسته را شناسایی کرده است.
- بیشترین اشتباه‌ها بین دسته‌های الکترونیک - لوازم خانگی و پوشاک - لوازم خانگی رخ داده‌اند.
- دسته‌ی الکترونیک بیشترین میزان اشتباه را دارد، که ممکن است به دلیل شباهت‌های متنی یا توصیفی با لوازم خانگی باشد.
- دسته‌ی کتاب‌ها همچنان با دقت بالا تشخیص داده می‌شود، اما چند مورد به اشتباه در دسته‌های دیگر قرار گرفته‌اند.

در کل TF-IDF نسبت به word2vec بهتر عمل کرده است. چون Word2Vec بیشتر اطلاعات معنایی را حمل می‌کند، اما به خوبی TF-IDF ویژگی‌های خاص متن را برای دسته‌بندی استخراج نمی‌کند (مگر اینکه context بسیار قوی باشد). Word2Vec در این مسئله خاص (با متن‌های کوتاه) کارایی کمتری نسبت به TF-IDF داشت.

برای متن‌هایی که معنی در ساختار جمله مهم‌تره (مثل مکالمه یا ترجمه)، Word2Vec ممکنه بهتر باشه.

الگوریتم‌های ساده‌تری مثل Logistic و SVC برای مسائل با داده زیاد و ویژگی زیاد (مثل TF-IDF) عملکرد بهتری دارن. TF-IDF گزینه‌ی بهتر و دقیق‌تری است، چون ویژگی‌های آماری سطح کلمه را بهتر استخراج می‌کند و برای الگوریتم‌هایی مثل SVM و Regression Logistic بهینه‌تر است. در حالی که Word2Vec برای کاربردهایی که درک معنایی و شباهت بین واژه‌ها اهمیت دارد، مناسب‌تر است، نه لزوماً طبقه‌بندی کلاسیک.

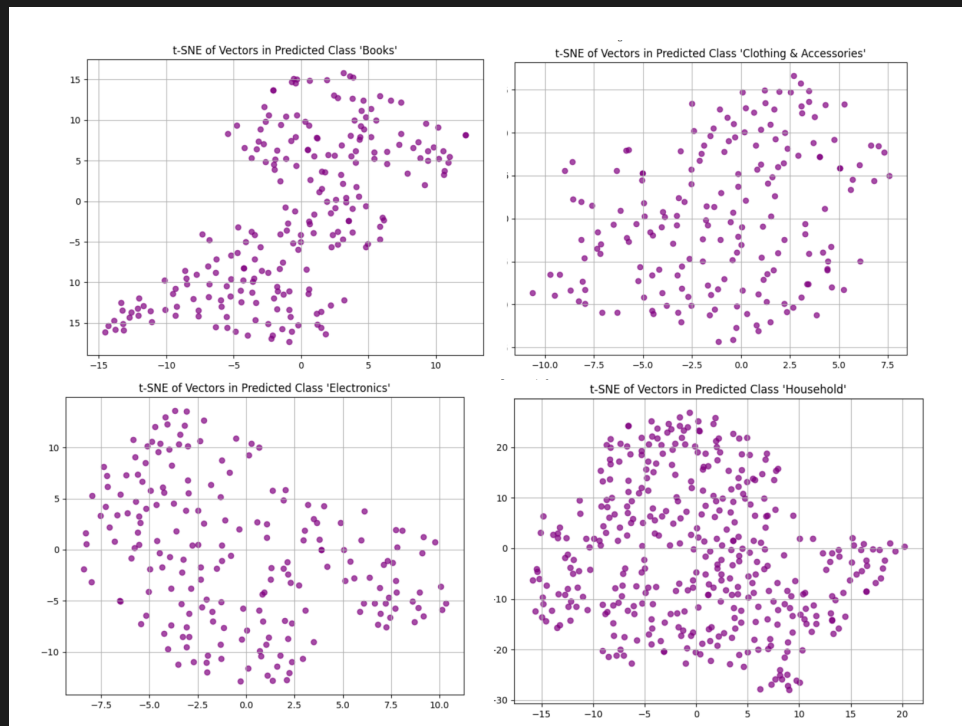
۳.۴ نشان دادن بردار های کلی



بردار های word2vec جداپذیر تر به نظر می رسد به صورت خوشه ای پراکنده شده اند کمی خطا هم دیده می شود که نسبت به تعداد داده ها قابل چشم پوشی است. بردار های tfidf به نظر شلوغ تر میرسد اما دقت بهتری دارد، که ممکن است به این دلیل باشد که برای نشان دادن داده ها بر روی ۲ بعد کاهش ابعاد انجام شده (ابعاد tfidf بسیار زیاد بوده است) و این کاهش ابعاد باعث شده است

که به این شکل دیده شود (یعنی در بعد های بالاتر جدا پذیر هستند) چون word2vec برحسب اطلاعات معنایی است از نزدیک بودن دسته ها به همدیگر می توان شباهت معنایی بین دسته ها را فهمید مثلا اینکه وسایل خانگی بین لباس ها و لوازم الکترونیک است و کتاب از بقیه دسته ها دور است.

۴.۴ نشان دادن بردار ها و چند مستند نمونه- tfidf



Predicted Class: Clothing & Accessories

Document 0

Text: Diverse Men's Formal Shirt Diverse is a western wear value brand for men. Our range consists of basic and updated basic apparel across both formal and casual wear. We offer the right blend of quality, style and value aimed to delight our customers.

True Label: Clothing & Accessories

Predicted Label: Clothing & Accessories

Document 1

Text: Jack's Star Soft Cotton Track Pants for Kids Infants & Toddler - Lowers/Joggers for Boys and Girls with Bottom Ribs- Pack of 5

True Label: Clothing & Accessories

Predicted Label: Clothing & Accessories

Document 9

Text: ROOLIUMS A, Brand Factory Outlet Women's Girls Stripe Tights for Yoga Gym and Active Sports Fitness

True Label: Clothing & Accessories

Predicted Label: Clothing & Accessories

Document 10

Text: AAKRITHI Women Formal Blazer Aakrithi comes with its trendy dynamic fitting women Blazer with its unmatched comfort level and unique colour combination which provide the feel of uniqueness and comfortness to its owner. Aakrithi have designed this women Blazer taking care the needs of its customer in mind. Aakrithi never compromises with the quality of products. Aakrithi is one of the biggest women Blazer Brand in India that provide high quality women Blazer at very Reasonable And affordable Cost.

True Label: Clothing & Accessories

Predicted Label: Clothing & Accessories

Document 12

Text: New Era Women's Polyester Taping Rain Coat (Black, XL) Treated to resist rain-and-stains, this sleek trench provides style, protection and comfort on chilly days.raincoat newera raincoat branded raincoat waterproof raincoat long rainwear for women/raincoat for women in rain coat for women/raincoats for women.

True Label: Clothing & Accessories

Predicted Label: Clothing & Accessories

Predicted Class: Books

Document 2

Text: UGC/NET/JRF/SET English Literature (Paper-II And III) UGC Pol Sci useful for DSSB KVS JNU etc

True Label: Books

Predicted Label: Books

Document 7

Text: Pharmacology for Dentistry About the Author Tara V Shanbhag is Professor and Head at the Department of Pharmacology, Srinivas Institute of Medical Sciences and

Research Centre, Mukka, Mangalore, Karnataka, India. She has a teaching experience of more than 30 years and has been an examiner to various universities. Dr Shanbhag is an honourable recipient of the 'Good Teacher' award. She has authored books on pharmacology and has to her credit several articles in national and international journals. Smita Shenoy is Additional Professor at the Department of Pharmacology, Kasturba Medical College, Manipal University, Manipal, Karnataka, India. She has 13 years of teaching experience and has been an examiner to various universities. Dr Shenoy has been honoured with the 'Good Teacher' award. She has authored books on pharmacology and has published several articles in national and international journals. Veena Nayak is Associate Professor at the Department of Pharmacology, Kasturba Medical College, Manipal University, Manipal, Karnataka, India. She has 10 years of teaching experience and has been an examiner to various universities. She is also a faculty and resource person at the Department of Medical Education, Kasturba Medical College, Manipal University, Manipal. She has authored books on pharmacology. Dr. Nayak has obtained ICMR grant for research and has published several articles in national and international journals.

True Label: Books

Predicted Label: Books

Document 8

Text: NOVICZ Skating Board Skate Board

True Label: Books

Predicted Label: Books

Document 13

Text: UP Police Constable : Practice Test Papers and Previous Papers (Solved) (Old Edition)

True Label: Books

Predicted Label: Books

Document 27

Text: English English Dictionary (Hb): English Word - Its Meaning In English Along with Sentence

True Label: Books

Predicted Label: Books

Predicted Class: Electronics

Document 3

Text: MAA-KU AC Axial Cooling Blower Exhaust Rotary Fan, Size : 4.75" inches (12 x 12 x 3.8 cm), Black

True Label: Household

Predicted Label: Electronics

Document 6

Text: Zoook Rocker Thunder 20 watts Bluetooth Speaker with Karaoke Mic/TF/FM/LED/USB/Party Speaker Zoook has upped its game again with the new rocker thunder. Rocker thunder brings together style, creativity and function. This portable sound machi is perfect for those summer time beach days. With the built-in Bluetooth capability, FM radio, aux-in, mic-in, usb and micro-SD inputs. You can enjoy music anyway want. Don't worry about a power outlet. The built-in rechargeable lithium battery allows you to listen to music for hours. Included microphone gets your karaoke nights started

True Label: Electronics

Predicted Label: Electronics

Document 15

Text: Myra® TouYinGer X7 Led Projector 1800 Lumens, 800*600 HDMI USB VGA TV Home Cinema, Support Red & Blue 3D Format Resolution: Native : 800x600 support 720p 1080p, Brightness: 1800 Lumens, Contrast ratio: 1000:1, Lens: F=126 mm(Manual focusing), Multimedia Interface: 1*VGA,1*USB,1*SD,1*HDMI,1*3.5mm Audio port,1*AV,1*TV, Lamp: LED lamp;20000 hours, Displayable colours: 16.7K, Projection Method: Front, Rear Ceiling Mount, Table top, Image Zoom: Electronic horizontal and vertical flip; image zoom, Keystone Correction: ±12 vertical, manual keystone correction, Projection Screen Size (inch): 37-130 inch, Projection Distance (m): 1.2-3.8 meter, Aspect Ratio: 16:9 Native, 4:3 compatible; Audio Formats: MP3,WMA, AAC, Video Formats: MPEG1,MPEG2, MPEG4, H264,RM,RMVB,MOV,MJPEG,VC1,DIVX,FLV, Picture Formats: JPEG ,BMP ,PNG, Power Supply: AC110V 240V 50Hz/60Hz, Technology Type: TFT- Single LCD Panel + LED Technology, Dimension: 212mm*150mm*78m Important Note For better result use in dark place no light, AC3 Audio not support, best result upto 80 inch, Not recommended for Office presentation and Educational use. USB support only Pendrive not hard disk and phone

True Label: Electronics

Predicted Label: Electronics

Document 18

Text: Canon Pixma G3000 All-in-One Wireless Ink Tank Colour Printer P-S-copy, 8.8ipm (mono), 5.0ipm (color), 4800dpi x 1200dpi, 6000 pages print black and 7000 pages print color with additional two balck ink bottle and WI-FI.

True Label: Electronics

Predicted Label: Electronics

Document 25

Text: Yamaha PSR-E-363 61-Key Touch Sensitive Portable Keyboard Size:E363 Touch-sensitive keys allow expressive dynamic control Play the keys heavily and you'll get louder tones, or play softly to achieve quieter sounds. The touch-sensitive keyboard will accurately reflect every nuance of your playing, making your performances musically expressive.

True Label: Electronics

Predicted Label: Electronics

Predicted Class: Household

Document 4

Text: Generic Super Soft Sheep Faux Fur Hairy Washable Pillow Chair Carpet Rugs Mat Specification: Ideal as a rug or draped across your favorite armchair. High quality floor carpets, also perfect for home decor. Suitable to use on tiles, wooden and laminate floors. Add a beautiful touch to any room in your home. Type: Rug Material: Imitation Wool Occasions: Bedroom, Dinning Room, Office, Living Room Features: Soft, Warm, Anti-Slip Size: 60cm x 90cm/23.62" x 35.43" (Approx.) Notes: Due to the light and screen setting difference, the item's color may be slightly different from the pictures. Please allow slight dimension difference due to different manual measurement. Package Includes: 1 x Rug

True Label: Household

Predicted Label: Household

Document 5

Text: AmazonBasics 16 Piece Wood Suit Hanger, Cherry Size:16 Piece Solid wood construction to hold your heaviest clothes;Designed with a chrome swivel hook;Precisely cut notches on each end allow for hanging straps;Product Dimensions: 17.4 x 0.5 x 9.4 inches (LxWxH);1-Year Limited Warranty.

True Label: Household

Predicted Label: Household

Document 11

Text: Coleman Instant Canopy Sunwall - Accessory Only Please Note: This is for the sidewall attachment ONLY. The 12' x 12' canopy is NOT included with the purchase of this item Get more shade and weather protection outdoors with the Coleman Sunwall Instant Canopy Accessory. Made with UV Guard™ coating on the canopy material for superior defense against the sun, its durable construction will also stand up to the wind. It's compatible with 12 x 12 ft. straight-leg Coleman® Instant Canopies (NOT included). The Coleman Sunwall Instant Canopy Accessory is ideal for sporting events, birthdays, picnics and much more. This canopy sunwall is easy to move and place where you need it most. It helps to protect you and your guests against the wind, rain and sun, and it also offers privacy. We are an Authorized Coleman Dealer!

True Label: Household

Predicted Label: Household

Document 14

Text: Ajanta Royal Drop-Down 3 - 5 x 7 Photo Frame (Brown Metallic) : A-55 Material - Premium Quality Of Synthetic Wood.It'S A Graceful Drop Down Frame For The Beauty Of Your Home Wall And Also Specially For The Columns(Pillars) For Interior.Photo Size - 3 - 5" X 7" Inch And Each Frame Outer Size 8.75" X 6.75" Inch, On Wall Aprox. 0.75 Feet X 2 Feet.Box Content : 1-Set Of 3 Photo Frame.It'S A Unique One With Exclusive Quality

True Label: Household

Predicted Label: Household

Document 16

Text: Joyful Studio Plastic Modular Drawer System, (Studio XL 4) Size:Studio XL 4 Multipurpose Storage System Studio Modular Drawer System from Joyful. The trendy colors add a touch of creative Elegance While organizing your stuff. It's use is only limited to your imagination Easy to assemble modular design Convenient to use Ample of space to organize in each drawer Great looks

True Label: Household

Predicted Label: Household

تحلیل:

- دسته‌بندی پیش‌بینی‌شده: کتاب‌ها (Books)
مدل در این دسته عملکرد قابل قبولی داشته و توانسته متون آموزشی، کتاب‌های آزمون، و منابع علمی را به‌درستی تشخیص دهد.
نمونه‌هایی مثل: Den- for Pharmacology و Literature English UGC/NET/JRF/SET

tistry به توضوح محتوای آموزشی دارند و مدل به درستی آن‌ها را در دسته‌ی کتاب‌ها قرار داده است.

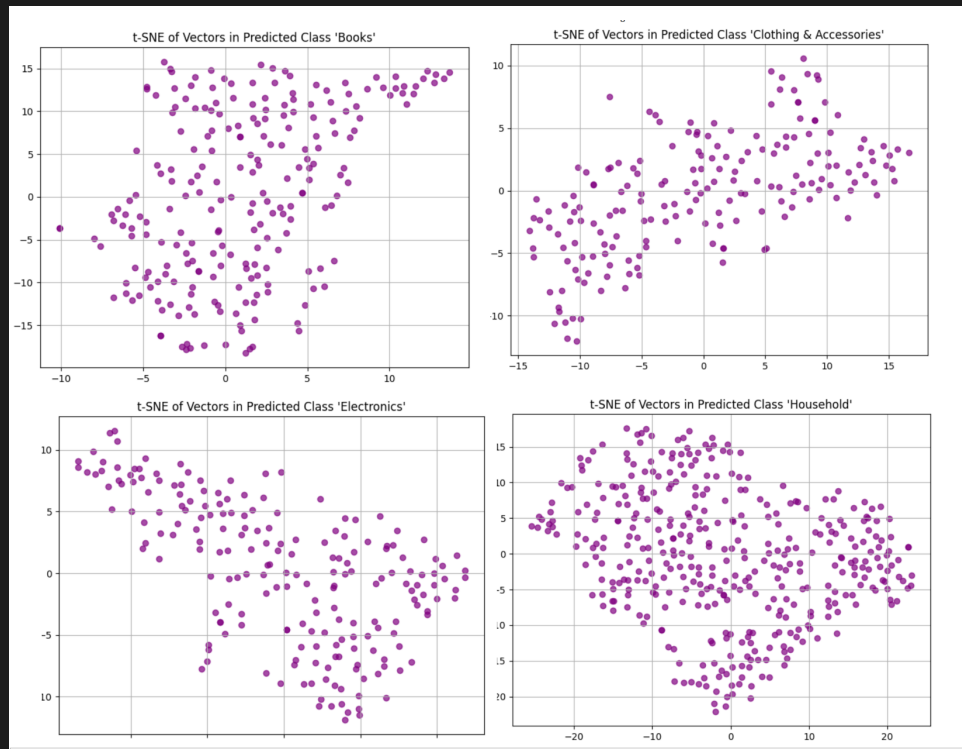
سند شماره ۸ با عنوان Board Skating NOVICZ به نظر کتاب نباشد. با اینکه مدل آن را به عنوان کتاب دسته‌بندی کرده، احتمالاً به دلیل وجود واژگان مشترک یا داده‌ی ناقص این اشتباه رخ داده است.

- دسته‌بندی پیش‌بینی‌شده: پوشاک و لوازم جانبی (Clothing & Accessories)
مدل در این دسته عملکرد بسیار خوبی داشته و همه‌ی نمونه‌ها را به درستی تشخیص داده است. نمونه‌هایی مثل: Rain و Blazer Women's Kids، for Pants Track Shirt، Formal Men's Coat همگی دارای توصیفاتی مرتبط با پوشاک هستند و مدل به درستی آن‌ها را دسته‌بندی کرده است.

- دسته‌بندی پیش‌بینی‌شده: الکترونیک (Electronics)
مدل در این دسته عملکرد نسبتاً خوب داشته، اما یک مورد اشتباه وجود دارد. نمونه‌هایی مثل: Portable و Printer Wireless Projector، LED Speaker، Bluetooth Keyboard همگی دارای مشخصات فنی و واژگان تخصصی مرتبط با الکترونیک هستند و مدل به درستی آن‌ها را تشخیص داده است.
سند شماره ۳ مربوط به Fan Blower Cooling است که برچسب واقعی آن لوازم خانگی بوده، اما مدل آن را به عنوان الکترونیک دسته‌بندی کرده. این اشتباه ممکن است به دلیل شباهت عملکردی یا واژگان مشترک باشد.

- دسته‌بندی پیش‌بینی‌شده: لوازم خانگی (Household)
مدل در این دسته نیز عملکرد خوبی داشته و همه‌ی نمونه‌ها را به درستی دسته‌بندی کرده است. نمونه‌هایی مثل: پادری خردار، چوب‌لباسی چوبی، دیوارپوش آفتاب‌گیر، قاب عکس، کشوی پلاستیکی همگی محصولات خانگی هستند و مدل به درستی آن‌ها را تشخیص داده است.

۵.۴ نشان دادن بردار ها و چند مستند نمونه - word2vec



Predicted Class: Books

Document 2

Text: UGC/NET/JRF/SET English Literature (Paper-II And III) UGC Pol Sci useful for DSSB KVS JNU etc

True Label: Books

Predicted Label: Books

Document 7

Text: Pharmacology for Dentistry About the Author Tara V Shanbhag is Professor and Head at the Department of Pharmacology, Srinivas Institute of Medical Sciences and Research Centre, Mukka, Mangalore, Karnataka, India. She has a teaching experience of more than 30 years and has been an examiner to various universities. Dr Shanbhag is an honourable recipient of the 'Good Teacher' award. She has authored books on pharmacology and has to her credit several articles in national and international journals. Smita Shenoy is Additional Professor at the Department of Pharmacology, Kasturba Medical College, Manipal University, Manipal, Karnataka, India. She has 13 years of teaching experience and has been an examiner to various universities. Dr Shenoy

has been honoured with the 'Good Teacher' award. She has authored books on pharmacology and has published several articles in national and international journals. Veena Nayak is Associate Professor at the Department of Pharmacology, Kasturba Medical College, Manipal University, Manipal, Karnataka, India. She has 10 years of teaching experience and has been an examiner to various universities. She is also a faculty and resource person at the Department of Medical Education, Kasturba Medical College, Manipal University, Manipal. She has authored books on pharmacology. Dr. Nayak has obtained ICMR grant for research and has published several articles in national and international journals.

True Label: Books

Predicted Label: Books

Document 8

Text: NOVICZ Skating Board Skate Board

True Label: Books

Predicted Label: Books

Document 13

Text: UP Police Constable : Practice Test Papers and Previous Papers (Solved) (Old Edition)

True Label: Books

Predicted Label: Books

Document 27

Text: English English Dictionary (Hb): English Word - Its Meaning In English Along with Sentence

True Label: Books

Predicted Label: Books

Predicted Class: Clothing & Accessorie

Document 0

Text: Diverse Men's Formal Shirt Diverse is a western wear value brand for men. Our range consists of basic and updated basic apparel across both formal and casual wear. We offer the right blend of quality, style and value aimed to delight our customers.

True Label: Clothing & Accessories

Predicted Label: Clothing & Accessories

Document 1

Text: Jack's Star Soft Cotton Track Pants for Kids Infants & Toddler - Lowers/Joggers for Boys and Girls with Bottom Ribs- Pack of 5

True Label: Clothing & Accessories

Predicted Label: Clothing & Accessories

Document 9

Text: ROOLIUMS A, Brand Factory Outlet Women's Girls Stripe Tights for Yoga Gym and Active Sports Fitness

True Label: Clothing & Accessories

Predicted Label: Clothing & Accessories

Document 10

Text: AAKRITHI Women Formal Blazer Aakrithi comes with its trendy dynamic fitting women Blazer with its unmatched comfort level and unique colour combination which provide the feel of uniqueness and comfortness to its owner. Aakrithi have designed this women Blazer taking care the needs of its customer in mind. Aakrithi never compromises with the quality of products. Aakrithi is one of the biggest women Blazer Brand in India that provide high quality women Blazer at very Reasonable And affordable Cost.

True Label: Clothing & Accessories

Predicted Label: Clothing & Accessories

Document 12

Text: New Era Women's Polyester Taping Rain Coat (Black, XL) Treated to resist rain-and-stains, this sleek trench provides style, protection and comfort on chilly days.raincoat newera raincoat branded raincoat waterproof raincoat long rainwear for women/raincoat for women in rain coat for women/raincoats for women.

True Label: Clothing & Accessories

Predicted Label: Clothing & Accessories

Predicted Class: Electronics

Document 3

Text: MAA-KU AC Axial Cooling Blower Exhaust Rotary Fan, Size : 4.75" inches (12 x

12 x 3.8 cm), Black
True Label: Household
Predicted Label: Electronics

Document 6

Text: Zoook Rocker Thunder 20 watts Bluetooth Speaker with Karaoke Mic/TF/FM/LED/USB/Party
Speaker Zoook has upped its game again with the new rocker thunder. Rocker thunder brings together style, creativity and function. This portable sound machi is perfect for those summer time beach days. With the built-in Bluetooth capability, FM radio, aux-in, mic-in, usb and micro-SD inputs. You can enjoy music anyway want. Don't worry about a power outlet. The built-in rechargeable lithium battery allows you to listen to music for hours. Included microphone gets your karaoke nights started
True Label: Electronics
Predicted Label: Electronics

Document 15

Text: Myra® TouYinGer X7 Led Projector 1800 Lumens, 800*600 HDMI USB VGA TV Home Cinema, Support Red & Blue 3D Format Resolution: Native : 800x600 support 720p 1080p, Brightness: 1800 Lumens, Contrast ratio: 1000:1, Lens: F=126 mm(Manual focusing), Multimedia Interface: 1*VGA,1*USB,1*SD,1*HDMI,1*3.5mm Audio port,1*AV,1*TV, Lamp: LED lamp;20000 hours, Displayable colours: 16.7K, Projection Method: Front, Rear Ceiling Mount, Table top, Image Zoom: Electronic horizontal and vertical flip; image zoom, Keystone Correction: ±12 vertical, manual keystone correction, Projection Screen Size (inch): 37-130 inch, Projection Distance (m): 1.2-3.8 meter, Aspect Ratio: 16:9 Native, 4:3 compatible; Audio Formats: MP3,WMA, AAC, Video Formats: MPEG1,MPEG2, MPEG4, H264,RM,RMVB,MOV,MJPEG,VC1,DIVX,FLV, Picture Formats: JPEG ,BMP ,PNG, Power Supply: AC110V 240V 50Hz/60Hz, Technology Type: TFT- Single LCD Panel + LED Technology, Dimension: 212mm*150mm*78mm
Important Note For better result use in dark place no light, AC3 Audio not support, best result upto 80 inch, Not recommended for Office presentation and Educational use. USB support only Pendrive not hard disk and phone
True Label: Electronics
Predicted Label: Electronics

Document 18

Text: Canon Pixma G3000 All-in-One Wireless Ink Tank Colour Printer P-S-copy, 8.8ipm (mono), 5.0ipm (color), 4800dpi x 1200dpi, 6000 pages print black and 7000 pages print color with additional two balck ink bottle and WI-Fi. True Label: Electronics

Predicted Label: Electronics

Document 25

Text: Yamaha PSR-E-363 61-Key Touch Sensitive Portable Keyboard Size:E363 Touch-sensitive keys allow expressive dynamic control Play the keys heavily and you'll get louder tones, or play softly to achieve quieter sounds. The touch-sensitive keyboard will accurately reflect every nuance of your playing, making your performances musically expressive.

True Label: Electronics

Predicted Label: Electronics

Predicted Class: Household

Document 4

Text: Generic Super Soft Sheep Faux Fur Hairy Washable Pillow Chair Carpet Rugs Mat Specification: Ideal as a rug or draped across your favorite armchair. High quality floor carpets, also perfect for home decor. Suitable to use on tiles, wooden and laminate floors. Add a beautiful touch to any room in your home. Type: Rug Material: Imitation Wool Occasions: Bedroom, Dining Room, Office, Living Room Features: Soft, Warm, Anti-Slip Size: 60cm x 90cm/23.62" x 35.43" (Approx.) Notes: Due to the light and screen setting difference, the item's color may be slightly different from the pictures. Please allow slight dimension difference due to different manual measurement. Package Includes: 1 x Rug

True Label: Household

Predicted Label: Household

Document 5

Text: AmazonBasics 16 Piece Wood Suit Hanger, Cherry Size:16 Piece Solid wood construction to hold your heaviest clothes;Designed with a chrome swivel hook;Precisely cut notches on each end allow for hanging straps;Product Dimensions: 17.4 x 0.5 x 9.4 inches (LxWxH);1-Year Limited Warranty.

True Label: Household

Predicted Label: Household

Document 11

Text: Coleman Instant Canopy Sunwall - Accessory Only Please Note: This is for the sidewall attachment ONLY. The 12' x 12' canopy is NOT included with the purchase of

this item Get more shade and weather protection outdoors with the Coleman Sunwall Instant Canopy Accessory. Made with UV Guard™ coating on the canopy material for superior defense against the sun, its durable construction will also stand up to the wind. It's compatible with 12 x 12 ft. straight-leg Coleman® Instant Canopies (NOT included). The Coleman Sunwall Instant Canopy Accessory is ideal for sporting events, birthdays, picnics and much more. This canopy sunwall is easy to move and place where you need it most. It helps to protect you and your guests against the wind, rain and sun, and it also offers privacy. We are an Authorized Coleman Dealer!

True Label: Household

Predicted Label: Household

Document 14

Text: Ajanta Royal Drop-Down 3 - 5 x 7 Photo Frame (Brown Metalic) : A-55 Material - Premium Quality Of Synthetic Wood.It'S A Graceful Drop Down Frame For The Beauty Of Your Home Wall And Also Specially For The Columns(Pillars) For Interior.Photo Size - 3 - 5" X 7" Inch And Each Frame Outer Size 8.75" X 6.75" Inch, On Wall Aprox. 0.75 Feet X 2 Feet.Box Content : 1-Set Of 3 Photo Frame.It'S A Unique One With Exclusive Quality

True Label: Household

Predicted Label: Household

Document 16

Text: Joyful Studio Plastic Modular Drawer System, (Studio XL 4) Size:Studio XL 4 Multipurpose Storage System Studio Modular Drawer System from Joyful. The trendy colors add a touch of creative Elegance While organizing your stuff. It's use is only limited to your imagination Easy to assemble modular design Convenient to use Ample of space to organize in each drawer Great looks

True Label: Household

Predicted Label: Household

تحلیل:

- دسته‌بندی پیش‌بینی‌شده: کتاب‌ها (Books)
مدل در این دسته عملکرد قابل قبولی داشته و توانسته متون آموزشی، کتاب‌های آزمون، و منابع علمی را به‌درستی تشخیص دهد.
نمونه‌هایی مثل: Den- for Pharmacology و Literature English UGC/NET/JRF/SET
tistry به‌وضوح محتوای آموزشی دارند و مدل به‌درستی آن‌ها را در دسته‌ی کتاب‌ها قرار داده است.
سند شماره ۸ با عنوان Board Skating NOVICZ به نظر کتاب نباشد. با اینکه مدل آن را به‌عنوان کتاب دسته‌بندی کرده، احتمالاً به دلیل وجود واژگان مشترک یا داده‌ی ناقص این اشتباه رخ داده

است. یا خطایی در برچسب‌گذاری داده‌های آموزشی باشد.

- دسته‌بندی پیش‌بینی‌شده: پوشاک و لوازم جانبی (Clothing & Accessories)
مدل در این دسته عملکرد بسیار خوبی داشته و همه‌ی نمونه‌ها را به‌درستی تشخیص داده است. نمونه‌هایی مثل: Rain و Blazer Women's، Kids for Pants Track، Shirt Formal Men's Coat همگی دارای توصیفاتی مرتبط با پوشاک هستند و مدل به‌درستی آن‌ها را دسته‌بندی کرده است.

- دسته‌بندی پیش‌بینی‌شده: الکترونیک (Electronics)
مدل در این دسته عملکرد نسبتاً خوب داشته، اما یک مورد اشتباه وجود دارد. نمونه‌هایی مثل: Portable و Printer Wireless، Projector LED، Speaker Bluetooth Keyboard همگی دارای مشخصات فنی و واژگان تخصصی مرتبط با الکترونیک هستند و مدل به‌درستی آن‌ها را تشخیص داده است.
سند شماره ۳ مربوط به Fan Blower Cooling است که برچسب واقعی آن «لوازم خانگی» بوده، اما مدل آن را به‌عنوان «الکترونیک» دسته‌بندی کرده. این اشتباه ممکن است به دلیل شباهت عملکردی یا واژگان مشترک بین فن‌های صنعتی و تجهیزات الکترونیکی رخ داده باشد.

- دسته‌بندی پیش‌بینی‌شده: لوازم خانگی (Household)
مدل در این دسته نیز عملکرد خوبی داشته و همه‌ی نمونه‌ها را به‌درستی دسته‌بندی کرده است. نمونه‌هایی مثل: پادری خردار، چوب‌لباسی چوبی، دیوارپوش آفتاب‌گیر، قاب عکس، و کشوی پلاستیکی همگی محصولات خانگی هستند و مدل به‌درستی آن‌ها را تشخیص داده است.