COMPUTATIONAL INTELLIGENCE

FINAL PROJECT DOCUMENTATION

Ferdowsi University of Mashhad
Department of Computer Engineering

SPRING 2025

نام و نام خانوادگی	شماره دانشجویی
امیرحسین افشار	۴۰۱۲۲۶۲۱۹۶
علیرضا صفار	۴۰۱۱۲۶۲۲۸۱

پیاده سازی پروژه در این رپو گیتهاب قابل مشاهده می باشد.

- https://github.com/AlirezaSaffar/ecommerce-text-classifier

۱) فاز صفرم: دیتا پروفایلینگ

در ابتدا و مانند هر پروژه ای که با یادگیری سر و کار دارد، دیتا پروفایلینگ را انجام دادیم که بتوانیم insight هایی در رابطه با دیتایی که بر روی آن کار میکنیم بدست بیاوریم.

۱. بررسی تعداد سطر های داده:

تعداد سطر های داده را بدست آوردیم که به شرح زیر است:

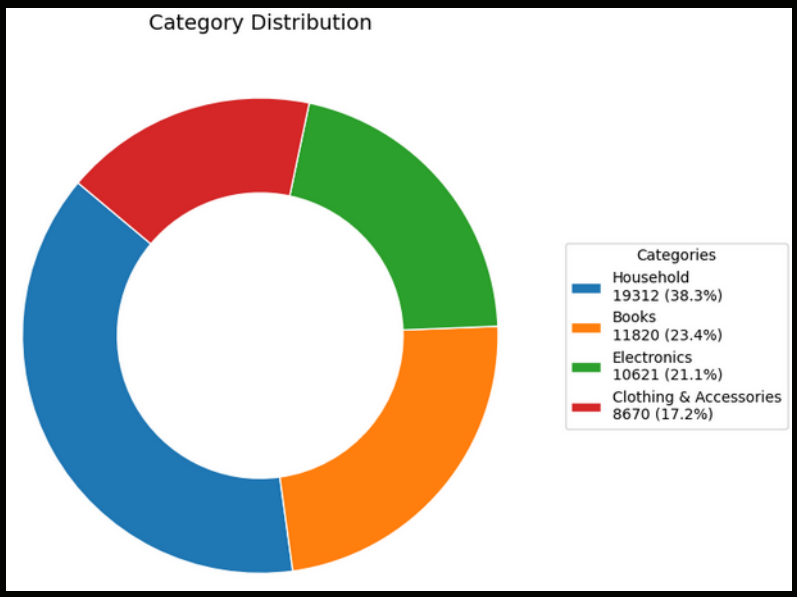
Property	Value
Shape	(۵۸۴۲۳، ۲)
Columns	’description’ ’category’

جدول ۱: اطلاعات کلی

۲. بررسی تعداد سطر های null و یا تکراری:

تعداد سطر های null برابر صفر بود، اما تعداد سطر های تکراری را برابر با مقدار تقریبی ۲۲k بدست آوردیم که تقریبا ۴۰ درصد دیتاست را تشکیل می داد. در فاز بعدی یعنی فاز اول: دیتا پریپروسسینگ، کل آنها را drop کردیم و فقط مقادیر unique را نگه داری کردیم. شایان ذکر است که در دیتاست اولیه، گاها حتی از یک دیتاپوینت بیش از ۳۰ بار تکرار داشتیم.

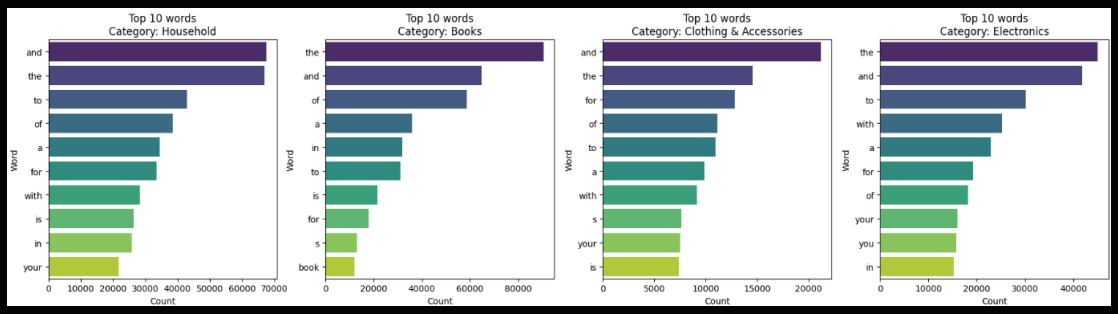
۳. بررسی تعداد کتگوری ها و میزان درصد هرکدام از آنها: همانطور که در شکل ۱ مشخص است، نزدیک به ۴۰ درصد داده ها را به تنهایی کتگوری household تشکیل داده اند و closing کمترین درصد را به خود اختصاص داده که نشان می دهد ممکن است در مرحله فاز اخر با بایاس شدن به سمت کتگوری ها رو به رو شویم. در این رابطه در بخش اخر بیشتر توضیح داده شده است.



شکل ۱: توزیع کتگوری ها

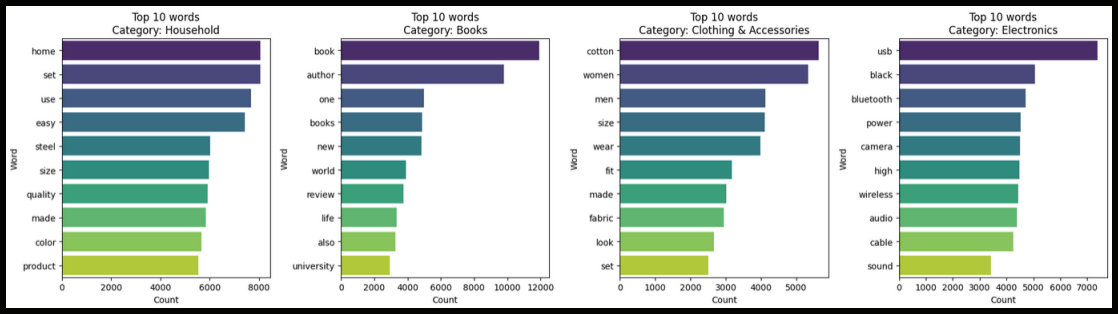
۴. بررسی کلمات پرتکرار هر کتگوری:

در ابتدا به شکل خام و سپس با اعمال حذف به شکل ساده این کار را انجام دادیم.



شکل ۲: نتیجه خام

همانطور که در شکل ۲ مشخص است نشان داده می شود که باید حذفیات کلمات غیرضروری اضافه صورت بگیرد تا بتوان به داده معناداری رسید.

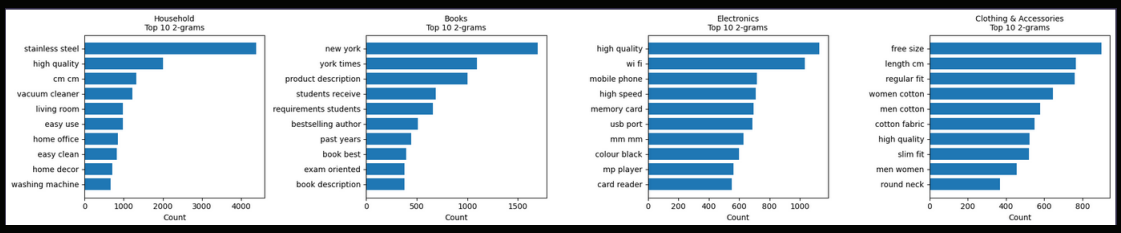


شکل ۳: نتیجه با اعمال حذف کلمات غیرضروری

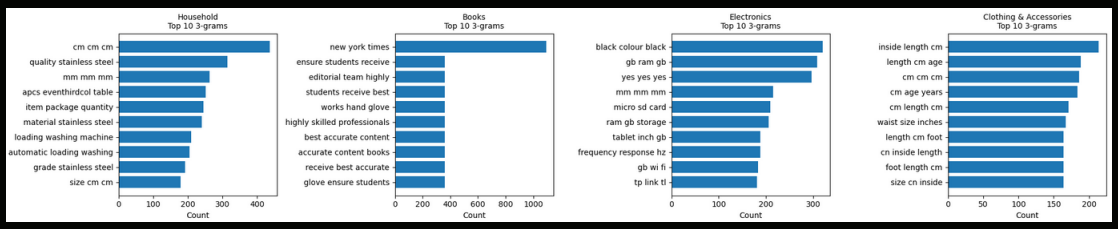
نتیجه شکل ۳ به طور کلی نشان می دهد که کلمات به خوبی در دامنه خود قابل تشخیص

هستند.

۵. بررسی n-gram ها به ازای ۲ و ۳:



شکل ۴: بررسی ۲گرم ها



شکل ۵: بررسی ۳گرم ها

شکل های ۴ و ۵ را بررسی کنید. با توجه به n-gram ها میتوان توجه ویژه ای به واحد ها و یونیت های اندازه گیری ای کرد. به این شکل که وقتی عبارتی نظیر:

a good quality table with 150 cm * 24 cm * 90 cm

رو به رو می شویم، پس از حذف ستاره و عدد ها با عبارت:

cm cm cm

رو به رو می شویم که به این صورت بیش از ۴۰۰ بار در دو کتگوری متفاوت رخ داده است و به طرز زیادی قرار است یادگیری بردارها را سخت کند. به این موضوع در پری پروسس توجه ویژه ای کردیم و عبارت بالا را کاملا درست کردیم و نرمالایز کردیم. توضیح بیشتر در بخش پری پروسس آمده است.

۲) فاز اول: پری پروسسینگ

در ابتدا سطر های تکراری را حذف کرده و سپس، برای پری پروسسینگ مراحل زیر را انجام دادیم و چنین پایپلاینی داشتیم:

Step	Function
0	Normalize units and remove singletons
1	Expand contractions
2	Convert to lowercase
3	Remove numbers
4	Remove punctuation
5	Remove special characters & emojis
6	Normalize whitespace
7	Tokenize text
8	Remove stopwords
9	Lemmatize tokens
10	Clean empty tokens

توضیحات بیشتر به شرح زیر است:

۱. Normalize-units-and-remove-singletons

همانطور که در بخش پروفایلینگ اشاره شد، واحدهای اندازه گیری مانند ،mhz gb، cm به شکل استاندارد تبدیل شدند و کاراکترهای تکراری حذف شدند و نرمالایز شدند.

۲. Expand-contractions

برای جملاتی که عموما به شکل not+verb خلاصه می شوند به کار بردیم.

۳. Convert-to-lowercase

برای این که همه کلمات یکنواخت باشند.

۴. Remove-numbers

از آنجا که اعداد نمی توانستند بردارهایی معنادار بسازند، همه اعداد را حذف کردیم

۵. Remove-punctuation

علائم نگارشی مانند کاما و نقطه برای تحلیل متن مفید نبودند و حذف شدند.

۶. Remove-special-characters-&-emojis

کاراکترهای خاص و ایموجی ها که معنای خاصی برای مدل نداشتند حذف شدند.

۷. Normalize-whitespace

فاصله های اضافی و تب ها به یک فاصله ساده تبدیل شدند.

۸. Tokenize-text

متن به کلمات جداگانه تقسیم شد تا قابل پردازش باشد.

۹. Remove-stopwords

کلمات رایج و بی معنی مانند ”the” و ”and” حذف شدند.

۱۰. Lemmatize-tokens

کلمات به شکل ریشه ای خود تبدیل شدند تا تنوع کاهش یابد.

۱۱. Clean-empty-tokens

توکن های خالی و بی معنی از نتیجه نهایی حذف شدند.

در نهایت یک مثال آورده می شود که اهمیت این پایپلاین دقیق تر نشان داده شود:
جمله ورودی:
SAF ’Floral’ Framed Painting (Wood, 30 inch x 10 inch, Special Effect UV Print Textured, SAO297) Painting made up in synthetic frame with UV textured print which gives multi effects and attracts towards it. This is an special series of paintings which makes your wall very beautiful and gives a royal touch (A perfect gift for your special ones).
و خروجی توکن های آن به این شکل در آمد:
[’saf’, ’floral’, ’frame’, ’paint’, ’wood’, ’numinch’, ’special’, ’effect’, ’uv’, ’print’, ’textured’, ’sao’, ’painting’, ’make’, ’synthetic’, ’frame’, ’uv’, ’textured’, ’print’, ’give’, ’multi’, ’effect’, ’attract’, ’towards’, ’special’, ’series’, ’painting’, ’make’, ’wall’, ’beautiful’, ’give’, ’royal’, ’touch’, ’perfect’, ’gift’, ’special’, ’one’]
مهم تر از همه توجهان را به بخش واحد های اندازه گیری جلب می کنیم که به جای
inch inch
به چنین توکن (بدون تکرار و یکبار آمده) تبدیل شده
numinch
و بنابراین کاملا هم ارتباط عدد و اندازه را حفظ می کند و هم اطلاعات با ارزشی را دور نمیریزد و هم نمایش بهتری را حاصل می شود.