دانشکده مهندسی کامپیوتر

استاد درس: سید صالح اعتمادی

بهار ۱۴۰۲

دسته‌بندی محتوای نامناسب در متن
پردازش زبان طبیعی

مرتضی شهرابی فراهانی
شماره دانشجویی: ۹۸۵۲۱۲۹۷

---

۱ مقدمه

با گسترش استفاده‌ی تلفن‌های هوشمند و فراگیر شدن شبکه‌های اجتماعی و سایت‌هایی که برای کاربران خود امکان نظردادن و یا صحبت کردن فراهم می‌کنند، امکان صحبت کردن و پیام فرستادن برای هر فردی که به این وسایل و نرم‌افزارها دسترسی دارد، فراهم شده است. اما در این بین، مشکلی که بوجود می‌آید، توهین‌ها، کلمات و جملات نامناسبی‌ست که امروزه و با استفاده از این دسترسی گسترده، هر کسی می‌تواند این موارد را به تعداد زیاد منتشر کند. باتوجه به زیاد بودن تعداد کاربران این شبکه‌های اجتماعی و پر تعداد بودن موارد نامناسبی از این قبیل، نیاز شد تا ابزارها و روش‌هایی ایجاد شوند تا در حد امکان، به صورت خودکار جلوی انتشار چنین پیام‌هایی را بگیرند.
هدف این پروژه، ایجاد ساز و کاری‌ست که به طور خودکار، موارد نامناسب در متن را تشخیص بدهد و در صورت استفاده در شبکه‌های اجتماعی، جلوی انتشار چنین پیام‌هایی را بگیرد و یا بعدا بتواند آن پیام‌ها را تشخیص دهد و اقدام به حذف آن‌ها بکند.
کد این پروژه در لینک زیر قرار داده شده است.
`https://github.com/morteza-shahrabi-farahani/NLP-project`
همچنین بعد از تغییرات لازم بر روی مجموعه دادگان پروژه، مجموعه دادگان جدید در huggingface با لینک زیر قرار داده شده است.
https://huggingface.co/datasets/Morteza-Shahrabi-Farahani/Detecting-toxic-comments

۲ مجموعه دادگان پروژه

در این پروژه، از مجموعه دادگان جمع‌آوری شده در یکی از مسابقات kaggle[۱] استفاده شده است. این مجموعه دادگان از کامنت‌های سایت wikipedia جمع‌آوری شده است. بخش train این مجموعه، شامل ۱۵۹۵۷۲ کامنت است که بخشی از این کامنت‌ها بدون محتوای نامناسب هستند و بخشی از آن‌ها دارای محتوای نامناسب هستند.
برای کامنت‌های نامناسب، ۶ دسته قرار داده شده است. این دسته‌ها شامل toxic, severe toxic, obscene, threat, insult, identity hate می‌باشند. معادل فارسی هر کدام از این دسته‌ها، به ترتیب برابر (سمی، خیلی سمی، مستهجن، تهدید، توهین و نفرت) می‌باشد. ازم به ذکر است، یک کامنت می‌تواند همزمان در چندین دسته از کامنت‌های محتوای نامناسب قرار داشته باشد. اما نمی‌تواند همزمان جزو کامنت‌های بدون مشکل باشد و عضو یکی از کلاس‌های بالا هم باشد.
هر کامت یک سطر از فایل csv. است. هر سطر شامل هشت ستون است. شش ستون همان دسته‌بندی‌های داده شده هستند که هر کدام از این ستون‌ها می‌تواند مقدار صفر یا یک داشته باشد. اگر مقدار آن ستون برابر یک بود، به معنای این است که این کامنت شامل آن دسته‌بندی می‌شود و در غیر این صورت و با داشتن مقدار صفر، کامنت جزو آن دسته قرار نمی‌گیرد. دو ستون دیگر، بیانگر آیدی کامنت و متن کامنت هستند.

۳ جمع‌آوری دادگان

با مراجعه به سایت kaggle و در قسمت مربوط به مجموعه دادگان مسابقه، داده‌های مسابقه قابل دریافت بود. این دیتاست شامل چند فایل با فرمت csv. می‌باشد. جهت راحت‌تر بودن دریافت دادگان، این چند فایل در

[۱] https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data

یکی از سایت‌های آپلود فایل قرار داده شدند و در ابتدای پروژه و در فایل collect data.py دریافت می‌شوند و در محیط پروژه ذخیره می‌شوند.
همچنین بعد از این مورد، لازم است که مجموعه دادگان هر کلاس را جدا کنیم. این کار در فایل classify data.py انجام شده و نتایج این بخش در محل data/raw/classified ذخیره شده است.

۴ تمیز کردن دادگان

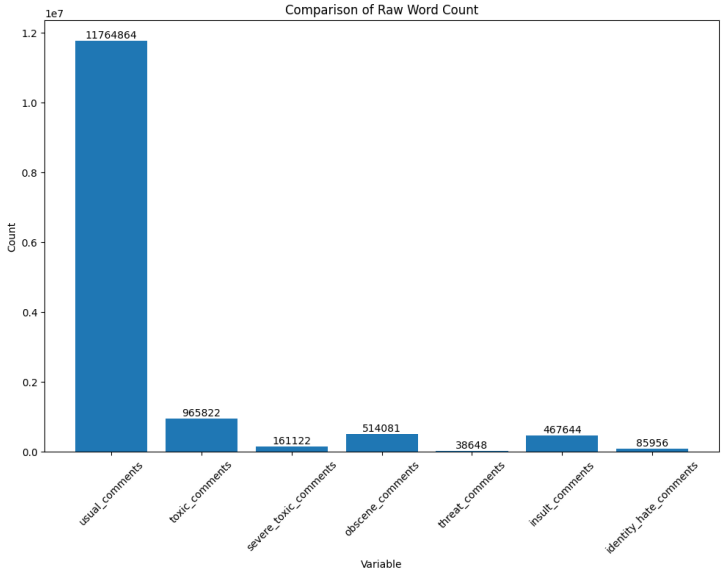
در این بخش، در ابتدا تمامی حروف تمامی کامنت‌ها به حروف کوچک تبدیل شدند. این کار موجب می‌شود تا تمایزی بین حروف بزرگ و کوچک قائل نشویم و برای مثال are و Are را یک کلمه حساب بکنیم. این کار موجب می‌شود تا تاثیر کلمات و مفاهیمشان بهتر توسط مدل درک شود.
متن‌های این مجموعه دادگان، به علت اینکه کامنت‌های سایت wikipedia بودند، شامل ایموجی و هشتگ و برخی کارکترهای مزاحم دیگر نبودند و یا تعدادشان کم بود. به همین علت نیاز به حذف کردن این موارد حس نشد.
همچنین امکان حذف کردن برخی موارد مانند علامات نگارشی و ... وجود داشت. اما به علت اینکه از کارکرد دقیق مدل مطلع نبودم، تصمیم به حذف کردن یا نکردن این موارد را به بعد از آموزش مدل موکول کردم. در هنگام آموزش مدل، در صورت بهتر شدن دقت مدل، می‌توان موارد این چنینی از جمله علائم نگارشی و یا برخی کلمات پرکاربرد اما بدون تاثیر را حذف کرد.
عملیات‌های مربوط به این بخش در فایل clean data.py انجام شده است و دادگان تمیز شده در پوشه‌ی data/clean ذخیره شده‌اند.

۵ آمار دادگان

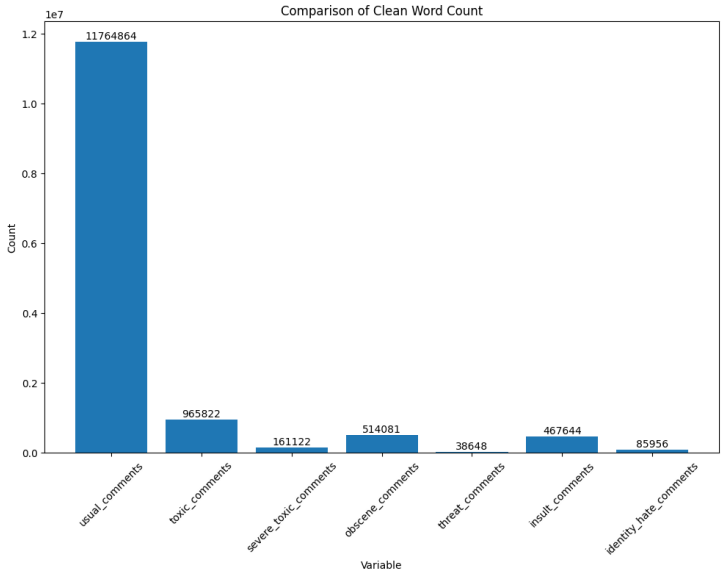
در این بخش برخی موارد آماری در مورد دادگان این پروژه بررسی شده‌اند. کدهای این بخش در فایل measure stats.py زده شده است.
برای ایجاد توکن از متن کامنت‌ها، از کتابخانه nltk استفاده شده است. همچنین برای تشخیص جملات هم از این کتابخانه استفاده شده است.
word tokenize: از این متد جهت تشخیص کلمات در هر کامنت استفاده شده است.
sent tokenize: از این متد جهت تشخیص و تمایز بین جملات استفاده شده است.
عملیات‌های مربوط به قسمت تشخیص کلمات و ملات کامنت‌ها در فایل tokenize classes.py انجام گرفته است.

۱.۵ مقایسه دادگان قبل و بعد از تمیزکردن

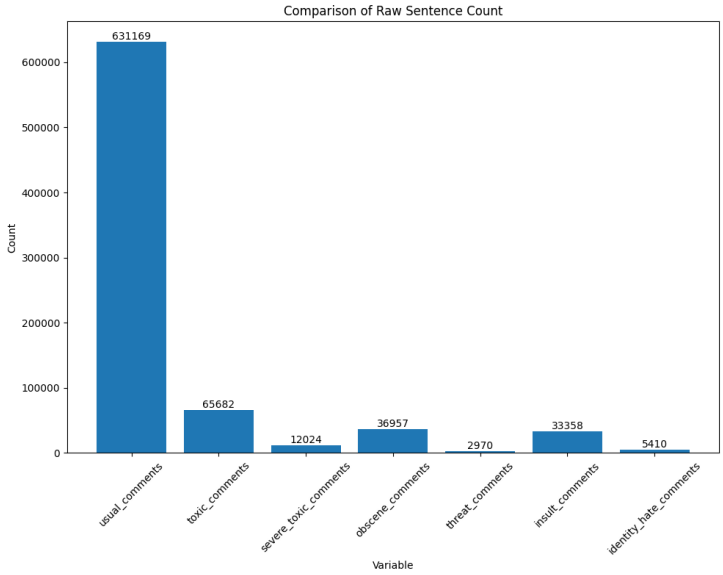
در نمودارهای زیر، تعداد کلمات و جملات قبل و بعد از تمیز کردن آن‌ها مقایسه و نشان داده شده است.



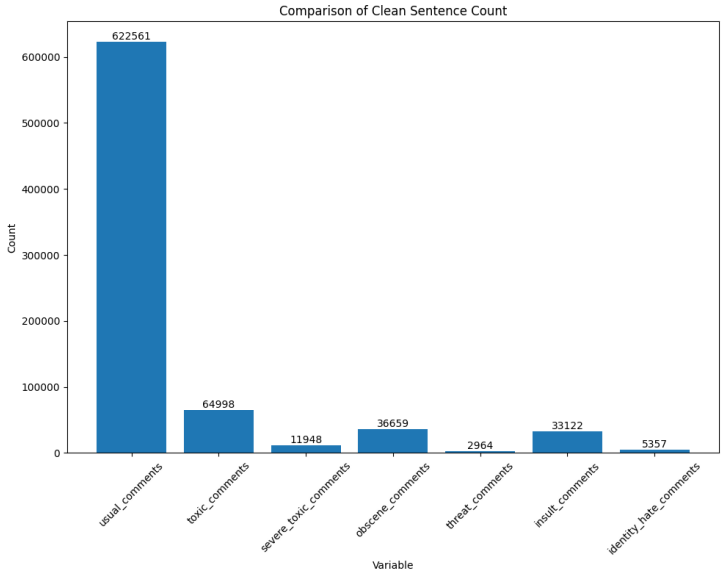
شکل ۱: تعداد کلمات قبل از تمیز کردن



شکل ۲: تعداد کلمات بعد از تمیز کردن



شکل ۳: تعداد جملات قبل از تمیز کردن



شکل ۴: تعداد جملات بعد از تمیز کردن

Category	Count
$\text{usual}_c omments$	11764864
$\text{toxic}_c omments$	965822
$\text{severe}_t oxic_c omments$	161122
$\text{obscene}_c omments$	514081
$\text{threat}_c omments$	38648
$\text{insult}_c omments$	467644
$\text{identity}_h ate_c omments$	85956

(a) Clean Words Count

Category	Count
$\text{usual}_c omments$	11771051
$\text{toxic}_c omments$	966192
$\text{severe}_t oxic_c omments$	161133
$\text{obscene}_c omments$	514198
$\text{threat}_c omments$	38652
$\text{insult}_c omments$	467731
$\text{identity}_h ate_c omments$	85985

(b) Raw Words Count

Category	Count
$\text{usual}_c omments$	622561
$\text{toxic}_c omments$	64998
$\text{severe}_t oxic_c omments$	11948
$\text{obscene}_c omments$	36659
$\text{threat}_c omments$	2964
$\text{insult}_c omments$	33122
$\text{identity}_h ate_c omments$	5357

(c) Clean Sentences Count

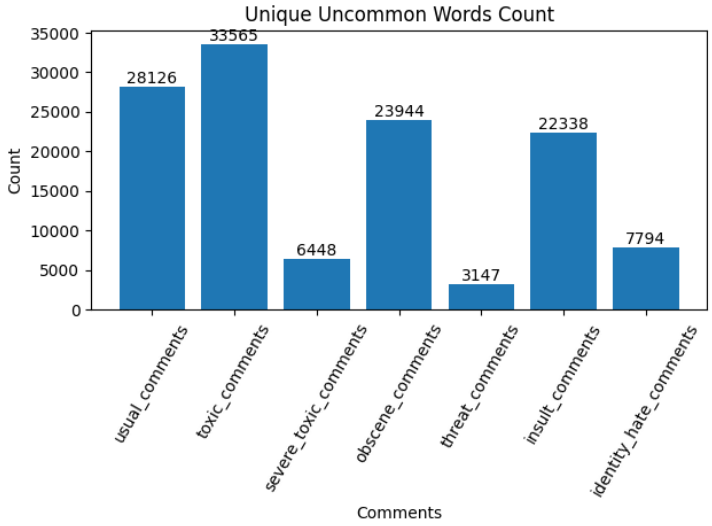
Category	Count
$\text{usual}_c omments$	631169
$\text{toxic}_c omments$	65682
$\text{severe}_t oxic_c omments$	12024
$\text{obscene}_c omments$	36957
$\text{threat}_c omments$	2970
$\text{insult}_c omments$	33358
$\text{identity}_h ate_c omments$	5410

(d) Raw Sentences Count

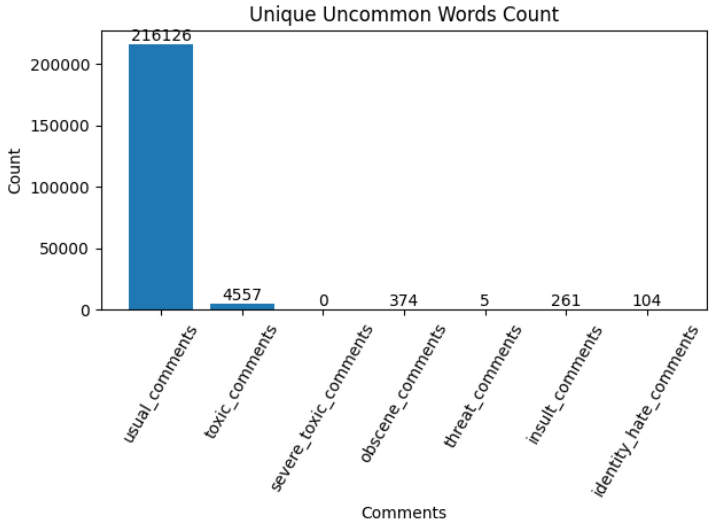
جدول ۱: جداول تعداد

۲.۵ بررسی تعداد داده‌های مشترک و غیر مشترک

نمودارهای این بخش، به صورت زیر هستند.



شکل ۵: تعداد کلمات مشترک هر دسته



شکل ۶: تعداد کلمات غیر مشترک هر دسته

Comment	Count
$\text{usual}_c\text{omments}$	28126
$\text{toxic}_c\text{omments}$	33565
$\text{severe}_t\text{oxic}_c\text{omments}$	6448
$\text{obscene}_c\text{omments}$	23944
$\text{threat}_c\text{omments}$	3147
$\text{insult}_c\text{omments}$	22338
$\text{identity}_h\text{ate}_c\text{omments}$	7794

(آ) تعداد کلمات مشترک در هر دسته

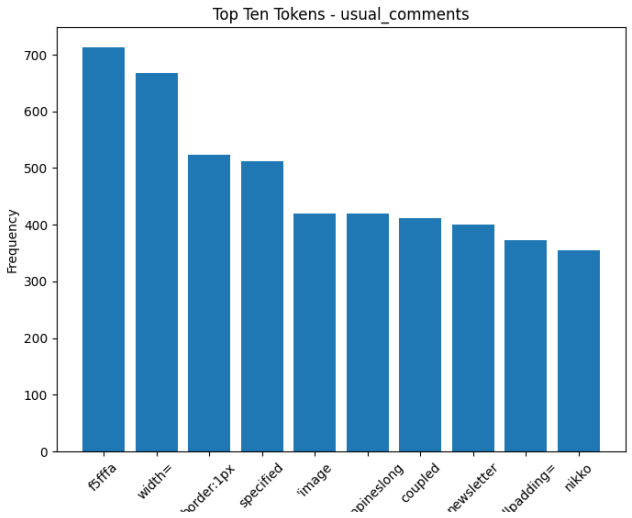
Comment	Count
$\text{usual}_c\text{omments}$	216126
$\text{toxic}_c\text{omments}$	4557
$\text{severe}_t\text{oxic}_c\text{omments}$	0
$\text{obscene}_c\text{omments}$	374
$\text{threat}_c\text{omments}$	5
$\text{insult}_c\text{omments}$	261
$\text{identity}_h\text{ate}_c\text{omments}$	104

(ب) تعداد کلمات غیر مشترک هر دسته

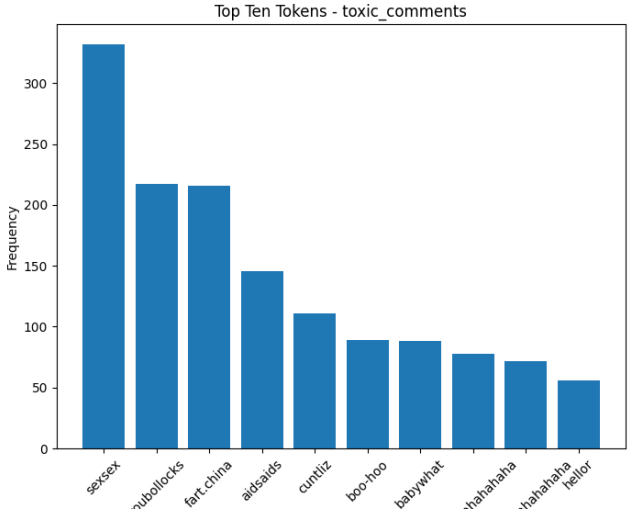
جدول ۲: جداول کلمات مشترک و غیر مشترک

۳.۵ بررسی کلمات غیر مشترک هر کلاس

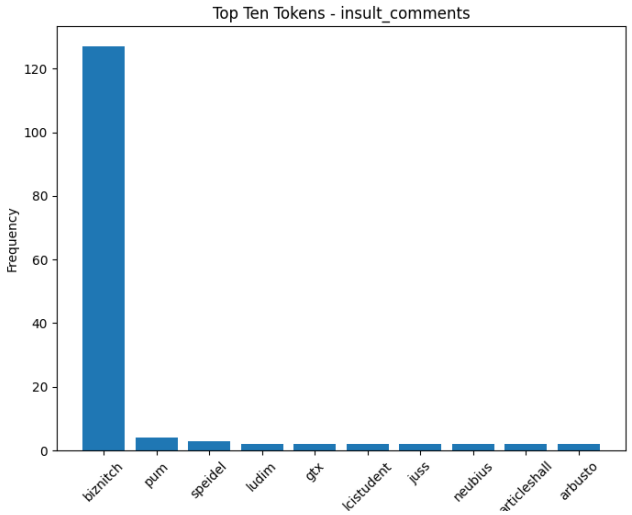
در نمودارها و تصاویر زیر، کلمات غیر مشترک پرتکرار هر کلاس آورده شده است. لازم به ذکر است کلاس severe toxic به علت اینکه درجه‌ی سختی بیشتری نسبت به سایر کلاس‌ها داشت، کلمه‌ی مستقلی که در سایر کلاس‌ها نیامده باشد را نداشت.



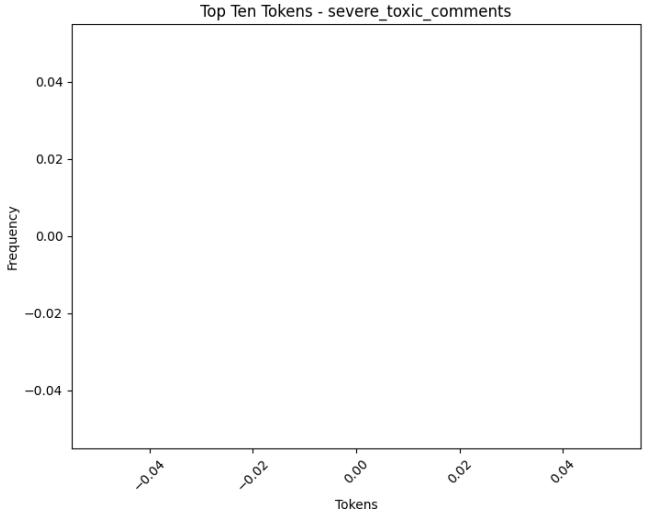
شکل ۷: ۱۰ کلمه غیر مشترک برتر جملات usual



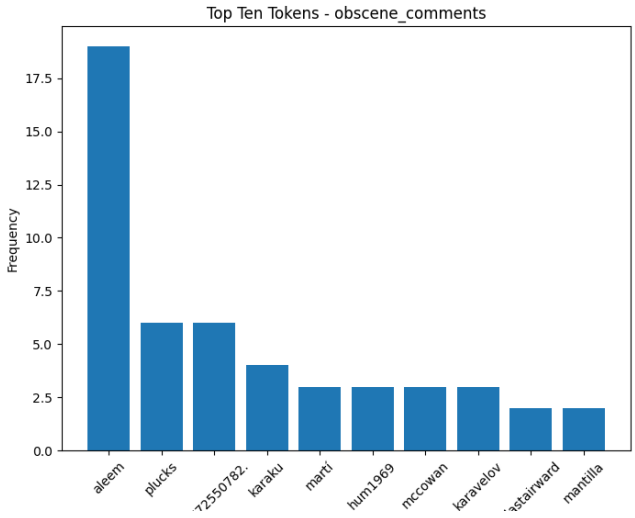
شکل ۸: ۱۰ کلمه غیر مشترک برتر جملات toxic



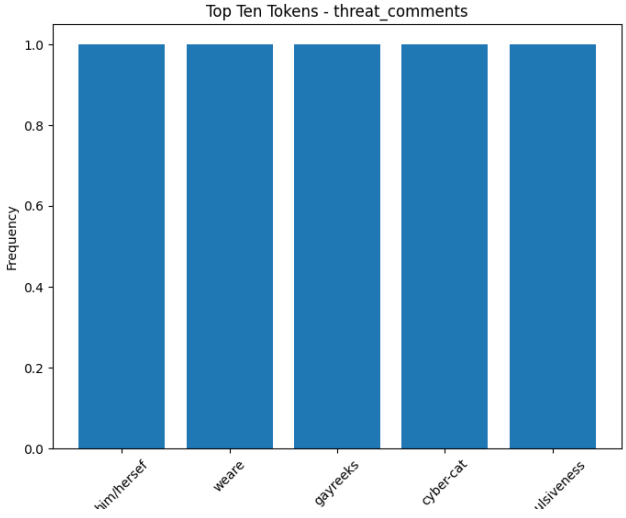
شکل ۹: ۱۰ کلمه غیر مشترک برتر جملات insult



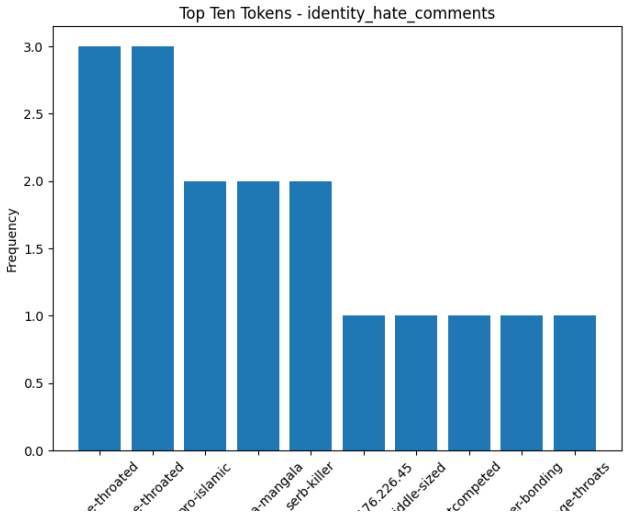
شکل ۱۰: ۱۰ کلمه غیر مشترک برتر جملات severe toxic



شکل ۱۱: ۱۰ کلمه غیر مشترک برتر جملات obscene



شکل ۱۲: ۱۰ کلمه غیر مشترک برتر جملات threat



شکل ۱۳: ۱۰ کلمه غیر مشترک برتر جملات identity hate

Token	Frequency
f5fffa	713
width=	668
border:1px	523
specified	512
'image	420
philippineslong	420
coupled	412
newsletter	400
cellpadding=	372
nikko	355

(آ) ۱۰ کلمه غیر مشترک برتر جملات usual

Token	Frequency
sexsex	332
youbollocks	217
fart.china	216
aidsaids	146
cuntliz	111
boo-hoo	89
babywhat	88
ahahahahahahaha	78
muahahahahahaha	72
hellor	56

(ب) ۱۰ کلمه غیر مشترک برتر جملات toxic

Token	Frequency
biznitch	127
pum	4
speidel	3
ludim	2
gtx	2
lcistudent	2
juss	2
neubius	2
articleshall	2
arbusto	2

(ج) ۱۰ کلمه غیر مشترک برتر جملات insult

Token	Frequency

(د) ۱۰ کلمه غیر مشترک برتر جملات severe toxic

Token	Frequency
aleem	19
plucks	6
07772550782.	6
karaku	4
mart	3
hum1969	3
mccowan	3
karavelov	3
alastairward	2
mantilla	2

(ه) ۱۰ کلمه غیر مشترک برتر جملات obscene

Token	Frequency
orange-throated	3
blue-throated	3
pro-islamic	2
chaitanya-mangala	2
serb-killer	2
longs124.176.226.45	1
middle-sized	1
outcompeted	1
stronger-bonding	1
orange-throats	1

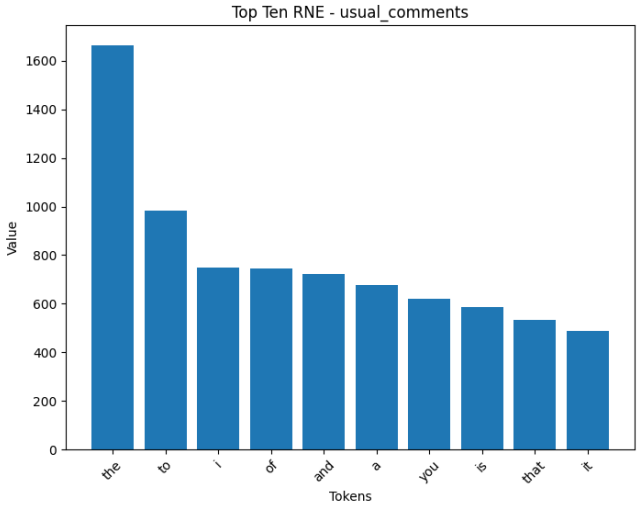
(و) ۱۰ کلمه غیر مشترک برتر جملات identity hate

Token	Frequency
him/hersef	1
weare	1
gayreeks	1
cyber-cat	1
repulsiveness	1

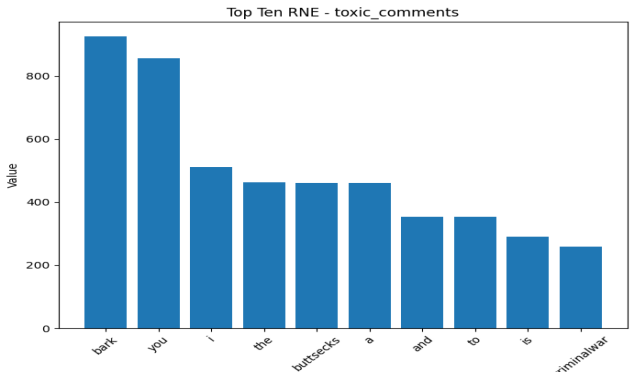
(ز) ۱۰ کلمه غیر مشترک برتر جملات threat

۴.۵ امتیازات کلمات هر دسته بر حسب RNE

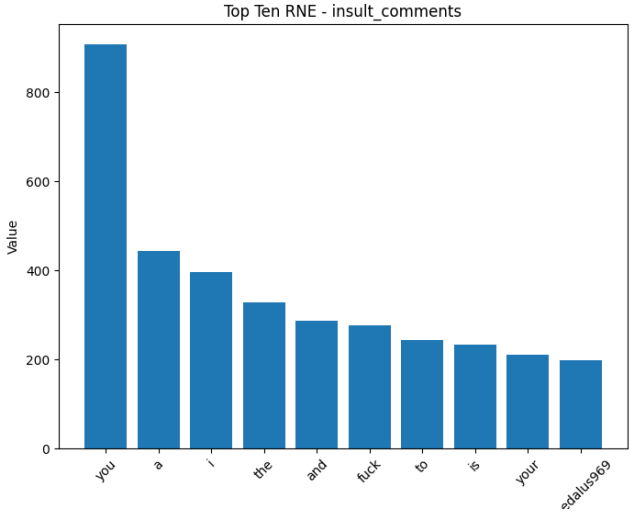
در این قسمت از گزارش، برای هر دسته، ۱۰ کلمه با بیشترین امتیاز طبق معیار توضیح داده شده در داکیومنت پروژه یعنی RNE آورده شده است.



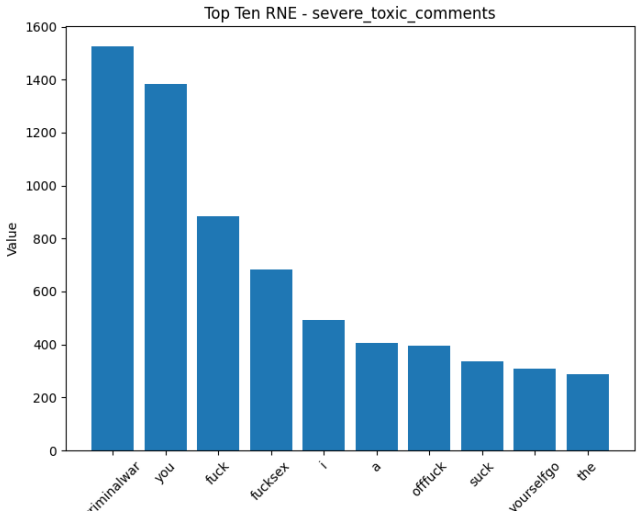
شکل ۱۴: ۱۰ کلمه برتر جملات usual



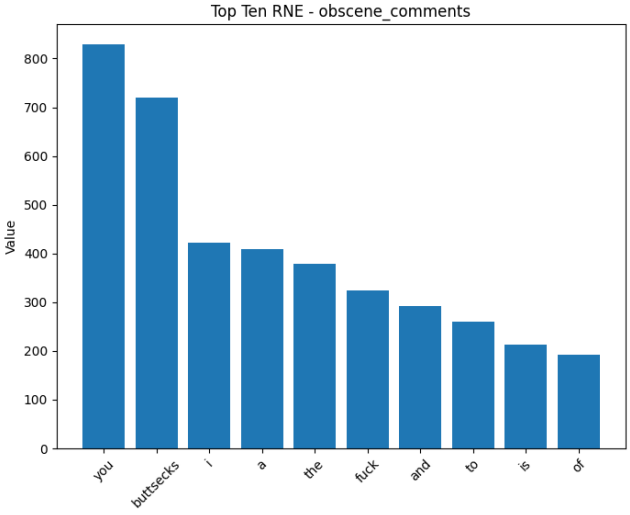
شکل ۱۵: ۱۰ کلمه برتر جملات toxic



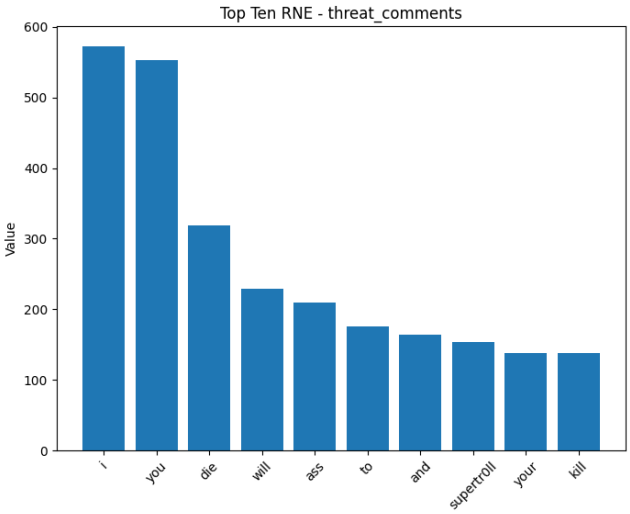
شکل ۱۶: ۱۰ کلمه برتر جملات insult



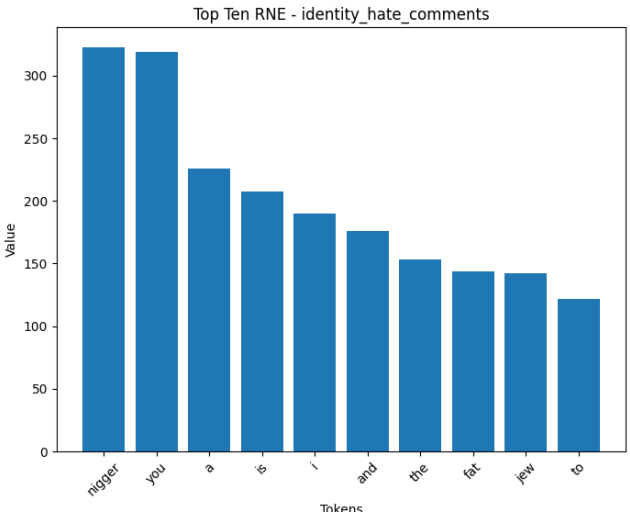
شکل ۱۷: ۱۰ کلمه برتر جملات severe toxic



شکل ۱۸: ۱۰ کلمه برتر جملات obscene



شکل ۱۹: ۱۰ کلمه برتر جملات threat



شکل ۲۰: ۱۰ کلمه برتر جملات identity hate

Token	Value
the	1663.5816
to	983.3859
i	747.9805
of	744.1572
and	722.8015
a	676.4604
you	621.8813
is	585.1654
that	532.6634
it	488.0157

(آ) ۱۰ کلمه برتر جملات usual

Token	Value
bark	925.7451
you	856.8122
i	511.1775
the	461.8688
buttsecks	460.5605
a	459.9515
and	354.3866
to	353.2137
is	290.0552
criminalwar	258.0249

(ب) ۱۰ کلمه برتر جملات toxic

Token	Value
you	907.7319
a	443.8587
i	395.8009
the	327.8808
and	286.9005
fuck	275.8278
to	243.1805
is	232.6405
your	209.7722
daedalus969	198.1288

(ج) ۱۰ کلمه برتر جملات insult

Token	Value
criminalwar	1525.5000
you	1384.2722
fuck	885.6408
fucksex	682.3742
i	492.0968
a	405.8131
offfuck	393.6774
suck	337.6664
yourselfgo	308.6789
the	286.5897

(د) ۱۰ کلمه برتر جملات severe toxic

Token	Value
you	829.2093
buttsecks	720.5458
i	422.5608
a	408.7701
the	377.8649
fuck	323.9397
and	291.6200
to	260.3258
is	213.1899
of	192.3271

(ه) ۱۰ کلمه برتر جملات obscene

Token	Value
nigger	322.7084
you	318.8977
a	225.9176
is	207.2999
i	189.6619
and	175.8347
the	152.9707
fat	143.7163
jew	142.3009
to	121.8322

(و) ۱۰ کلمه برتر جملات identity hate

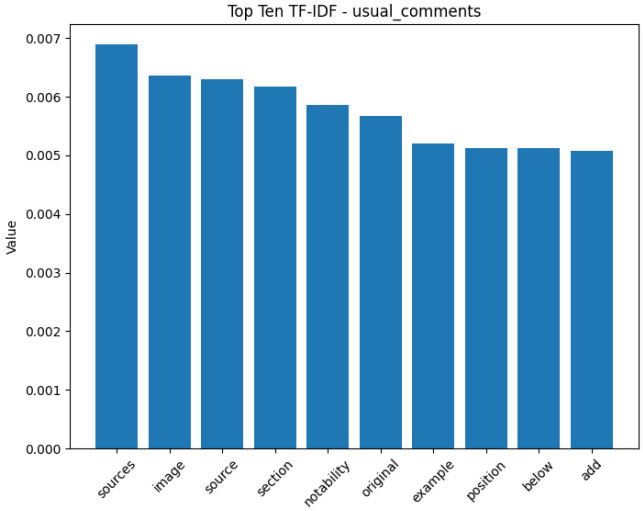
Token	Value
i	572.6317
you	553.5349
die	318.3712
will	229.1618
ass	209.7922
to	175.9635
and	163.9598
supertr0ll	153.5434
your	137.7699
kill	137.4971

(ز) ۱۰ کلمه برتر جملات threat

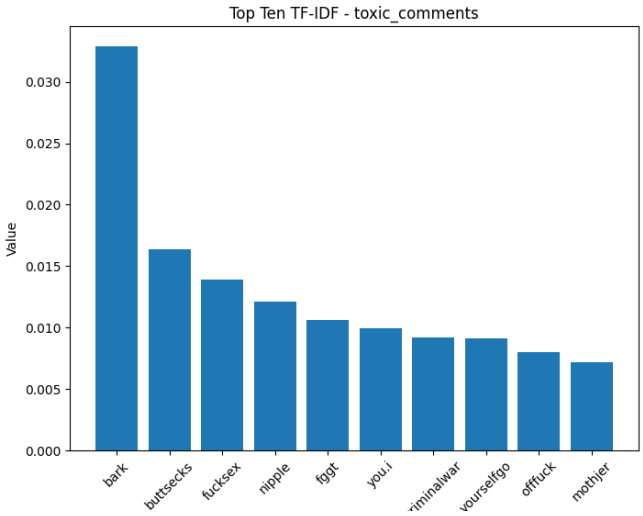
جدوال مربوط به امتیازات RNE در هر دسته

۵.۵ امتیازات کلمات هر دسته بر حسب TF-IDF

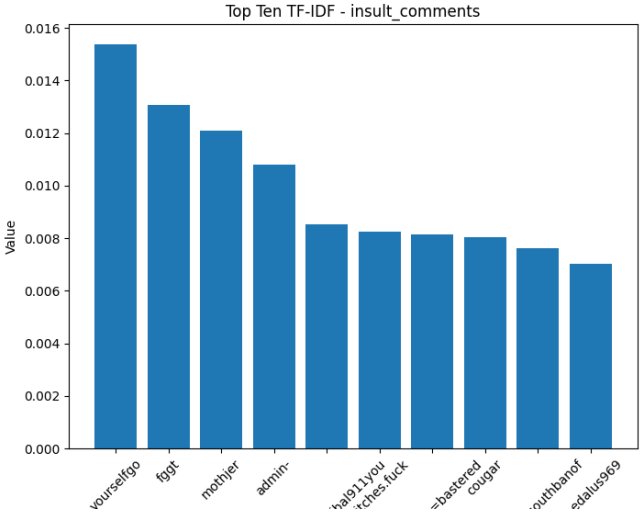
در این قسمت از گزارش، برای هر دسته، ۱۰ کلمه با بیشترین امتیاز طبق معیار توضیح داده شده در داکیومنت پروژه یعنی TF-IDF آورده شده است.



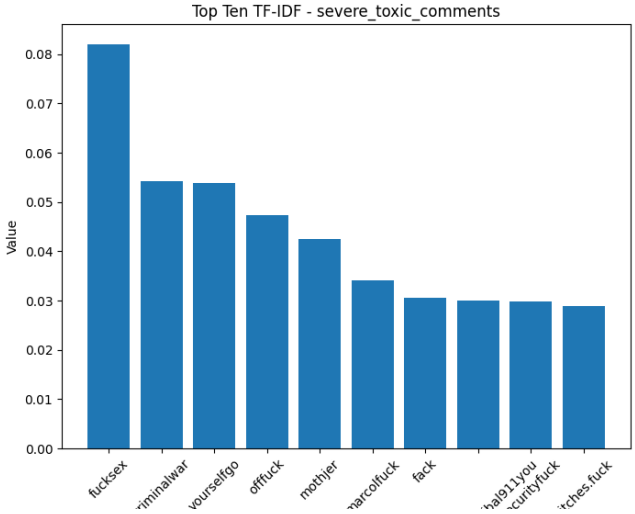
شکل ۲۱: ۱۰ کلمه برتر جملات usual



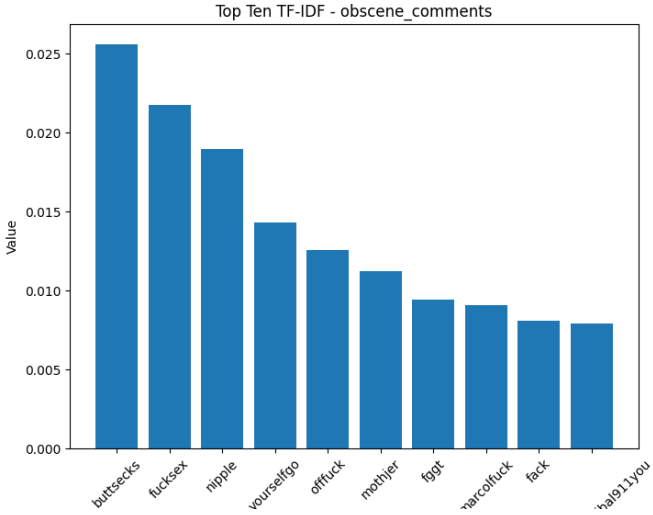
شکل ۲۲: ۱۰ کلمه برتر جملات toxic



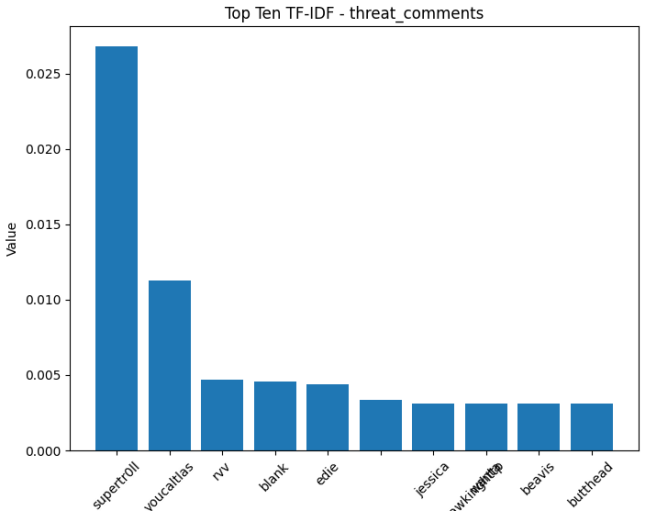
شکل ۲۳: ۱۰ کلمه برتر جملات insult



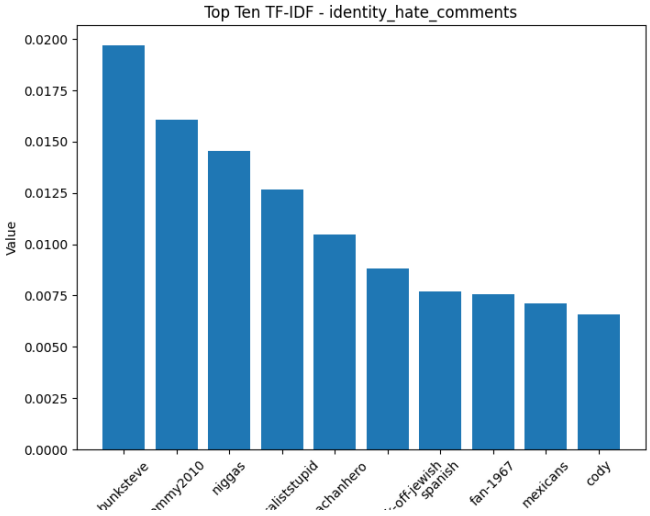
شکل ۲۴: ۱۰ کلمه برتر جملات severe toxic



شکل ۲۵: ۱۰ کلمه برتر جملات obscene



شکل ۲۶: ۱۰ کلمه برتر جملات threat



شکل ۲۷: ۱۰ کلمه برتر جملات identity hate

Token	Value
sources	0.00689302
image	0.00636225
source	0.00629535
section	0.00617797
notability	0.00585157
original	0.00567004
example	0.00520030
position	0.00511841
below	0.00511841
add	0.00507226

(آ) ۱۰ کلمه برتر جملات usual

Token	Value
bark	0.03289480
buttsecks	0.01636525
fucksex	0.01386900
nipple	0.01211067
fggt	0.01060178
you.i	0.00997945
criminalwar	0.00916848
yourselfgo	0.00911603
offfuck	0.00800134
mothjer	0.00717833

(ب) ۱۰ کلمه برتر جملات toxic

Token	Value
yourselfgo	0.01537773
fggt	0.01308496
mothjer	0.01210904
admin-	0.01079790
pro-assad.hanibal911you	0.00854319
bitches.fuck	0.00824603
bastered==bastered	0.00813592
cougar	0.00802344
notrhbysouthbanof	0.00762696
daedalus969	0.00704017

(ج) ۱۰ کلمه برتر جملات insult

Token	Value
fucksex	0.08199657
criminalwar	0.05420609
yourselfgo	0.05389600
offfuck	0.04730571
mothjer	0.04243984
marcolfuck	0.03416524
fack	0.03048590
pro-assad.hanibal911you	0.02994222
securityfuck	0.02982888
bitches.fuck	0.02890075

(د) ۱۰ کلمه برتر جملات severe toxic

Token	Value
buttsecks	0.02560339
fucksex	0.02174167
nipple	0.01896223
yourselfgo	0.01429071
offfuck	0.01254327
mothjer	0.01125307
fggt	0.00940745
marcolfuck	0.00905903
fack	0.00808344
pro-assad.hanibal911you	0.00793928

(ه) ۱۰ کلمه برتر جملات obscene

Token	Value
bunksteve	0.01969780
tommy2010	0.01608417
niggas	0.01456996
centraliststupid	0.01268311
bleachanhero	0.01048660
ancestryfuck-off-jewish	0.00881866
spanish	0.00771100
fan-1967	0.00758319
mexicans	0.00712396
cody	0.00658955

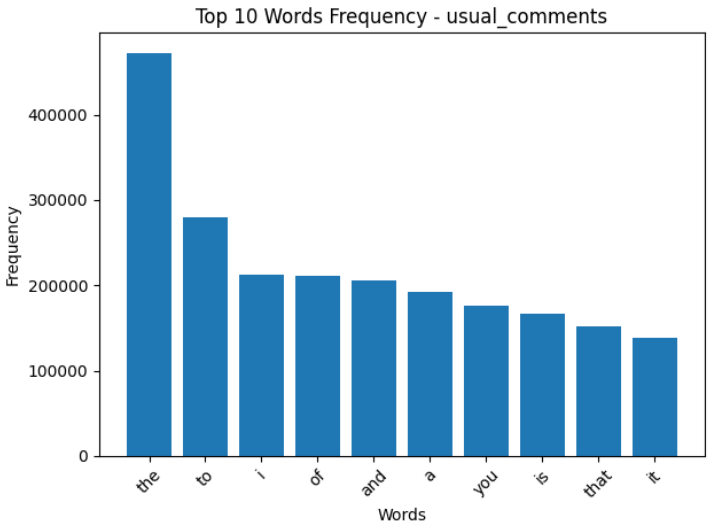
(و) ۱۰ کلمه برتر جملات identity hate

Token	Value
supertr0ll	0.02680901
youcaltlas	0.01129014
rvv	0.00469494
blank	0.00459713
edie	0.00440151
//en.wikipedia.org/wiki	0.00337449
jessica	0.00309572
wanta	0.00309572
beavis	0.00309572
butthead	0.00309572

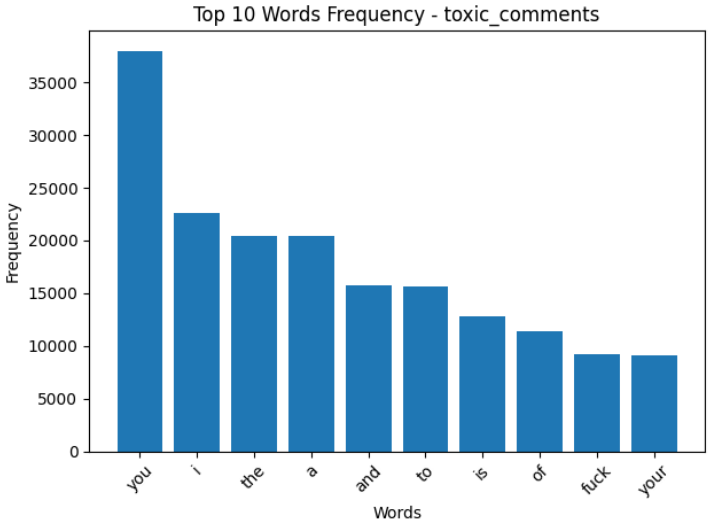
(ز) ۱۰ کلمه برتر جملات threat

جدوال مربوط به امتیازات TF-IDF در هر دسته

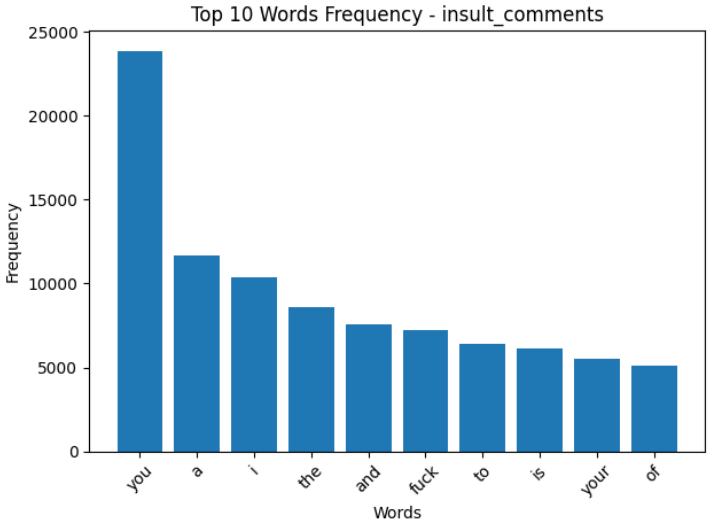
۶.۵ هیستوگرام کلمات پر تکرار



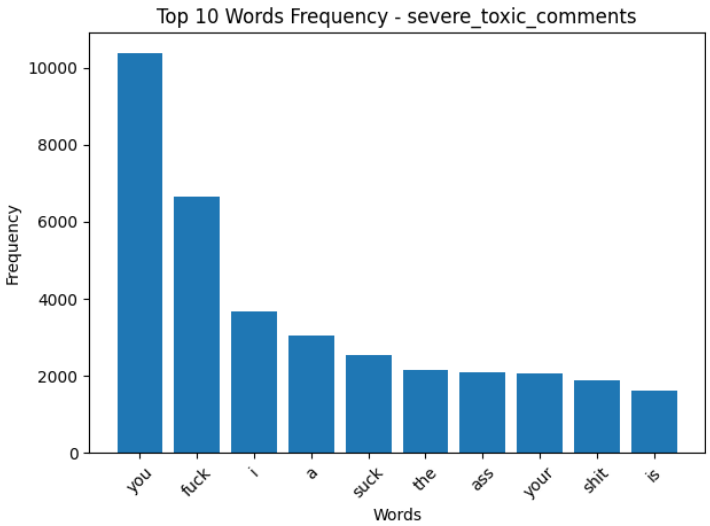
شکل ۲۸: ۱۰ کلمه برتر جملات usual



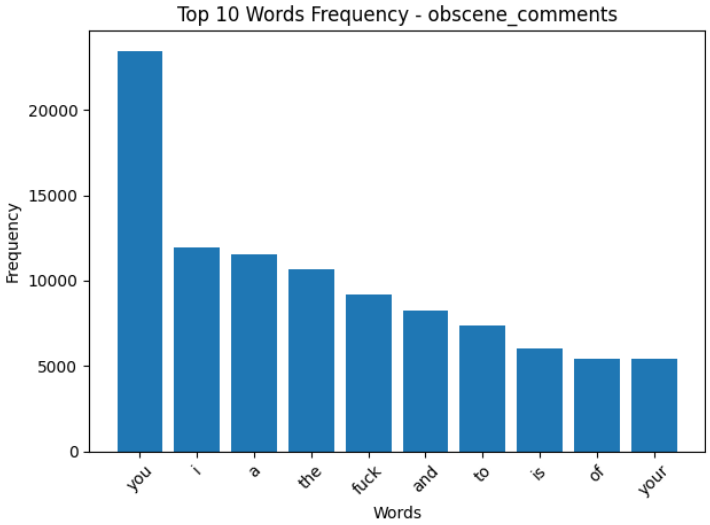
شکل ۲۹: ۱۰ کلمه برتر جملات toxic



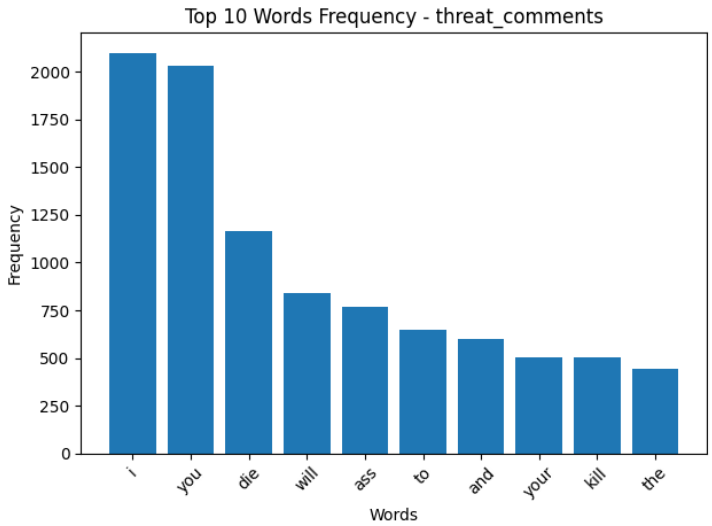
شکل ۳۰: ۱۰ کلمه برتر جملات insult



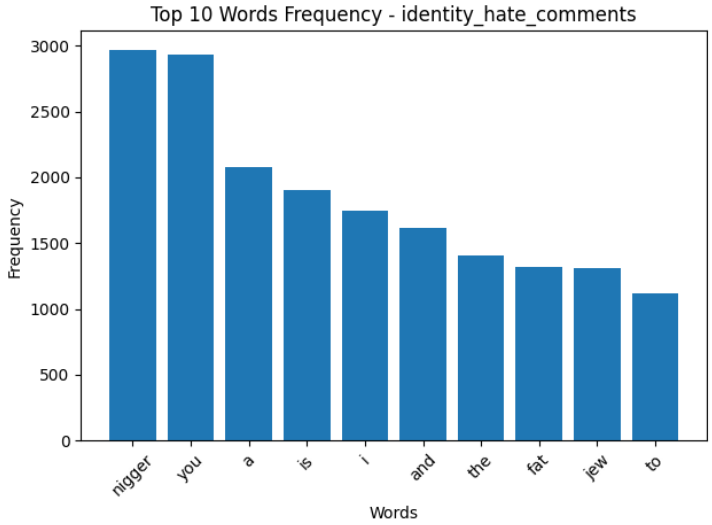
شکل ۳۱: ۱۰ کلمه برتر جملات severe toxic



شکل ۳۲: ۱۰ کلمه برتر جملات obscene



شکل ۳۳: ۱۰ کلمه برتر جملات threat



شکل ۳۴: ۱۰ کلمه برتر جملات identity hate

Token	Value
the	472534
to	279327
i	212461
of	211375
and	205309
a	192146
you	176643
is	166214
that	151301
it	138619

(آ) ۱۰ کلمه برتر جملات usual

Token	Value
you	37985
i	22662
the	20476
a	20391
and	15711
to	15659
is	12859
of	11387
fuck	9238
your	9125

(ب) ۱۰ کلمه برتر جملات toxic

Token	Value
you	23856
a	11665
i	10402
the	8617
and	7540
fuck	7249
to	6391
is	6114
your	5513
of	5079

(ج) ۱۰ کلمه برتر جملات insult

Token	Value
you	10380
fuck	6641
i	3690
a	3043
suck	2532
the	2149
ass	2105
your	2056
shit	1878
is	1631

(د) ۱۰ کلمه برتر جملات severe toxic

Token	Value
you	23450
i	11950
a	11560
the	10686
fuck	9161
and	8247
to	7362
is	6029
of	5439
your	5421

(ه) ۱۰ کلمه برتر جملات obscene

Token	Value
nigger	2964
you	2929
a	2075
is	1904
i	1742
and	1615
the	1405
fat	1320
jew	1307
to	1119

(و) ۱۰ کلمه برتر جملات identity hate

Token	Value
i	2099
you	2029
die	1167
will	840
ass	769
to	645
and	601
your	505
kill	504
the	445

(ز) ۱۰ کلمه برتر جملات threat

جدوال مربوط به امتیازات histogram در هر دسته