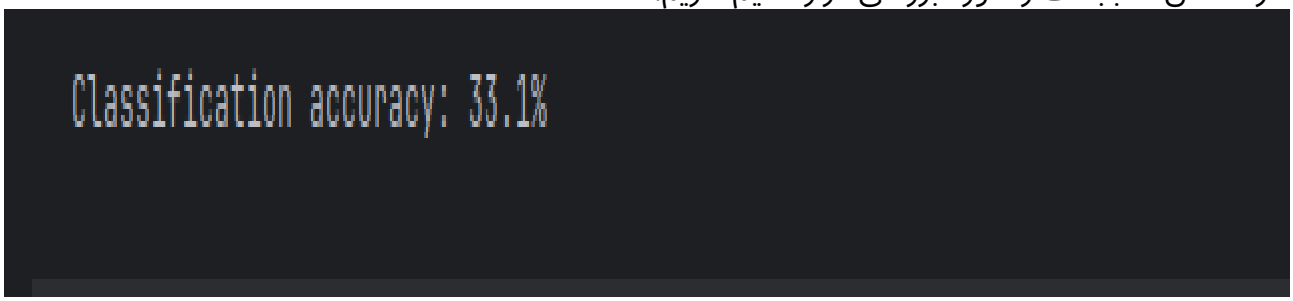
به نام خدا

محمد حسین رنگ‌رز – ۹۹۳۶۱۳۰۳۱

1.a:

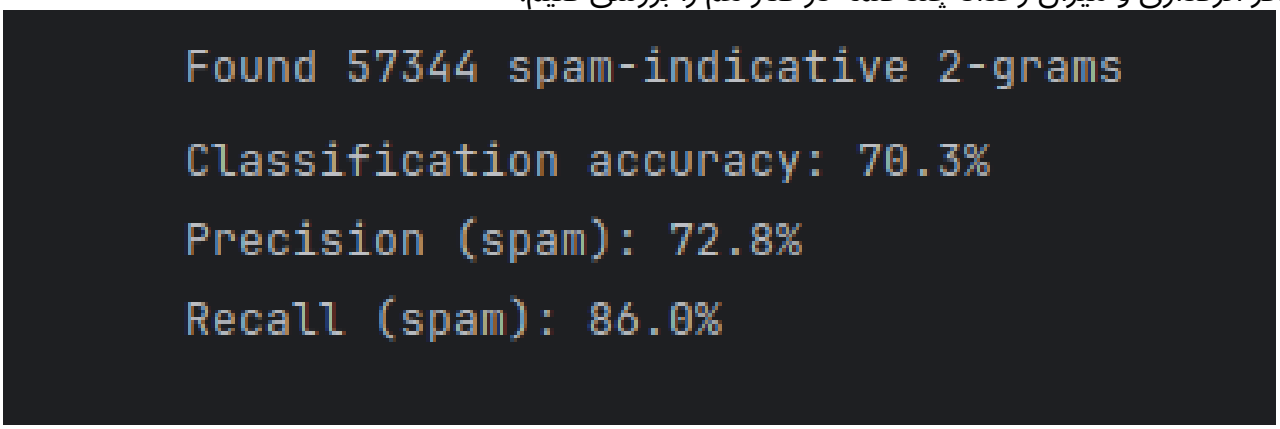
در این قسمت قرار است با متریک های مختلف اثرگذاری را بررسی کنیم.

اگر داشتن سابجکت را مورد بررسی قرار دهیم داریم:



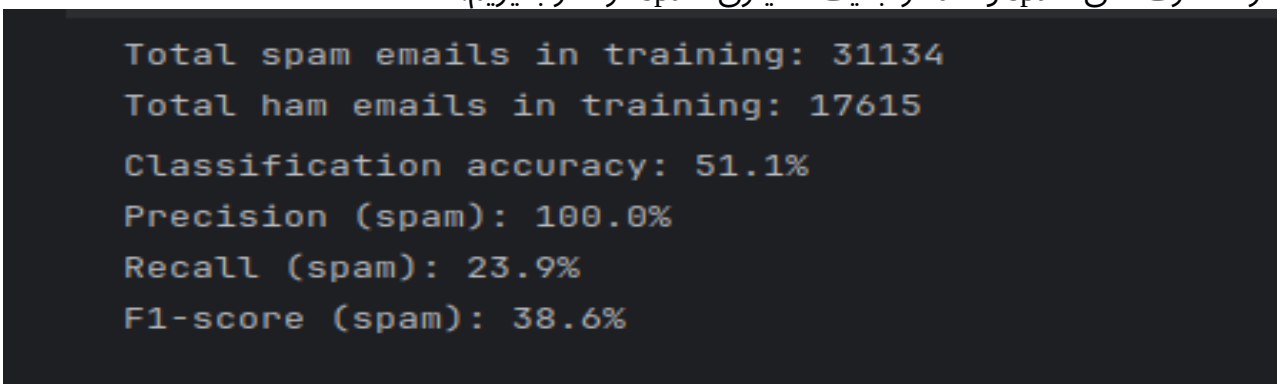

میتوان مشاهده کرد که دقت به شدت بدی به ما داد و بررسی سابجکت نمیتواند یک متریک خوبی برای پیدا کردن اسپم باشد و فایده‌ای ندارد.

اگر اثرگذاری و میزان رخداد چند کلمه در کنار هم را بررسی کنیم:

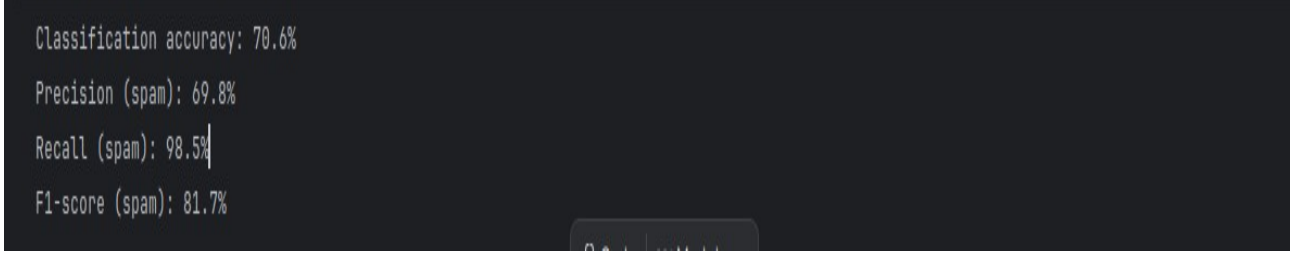



باز هم میتوان مشاهده کرد که تأثیر خوبی برروی پیدا کردن اسپم نداشته است که دقت به حد خیلی خوبی افزایش پیدا کند البته میتوان با N بیشتر بررسی کرد در این مثال صرفاً ۲ گرام بررسی شده است.

اگر مشترک های spam و ham را با یک امتیازی spam در نظر بگیریم:

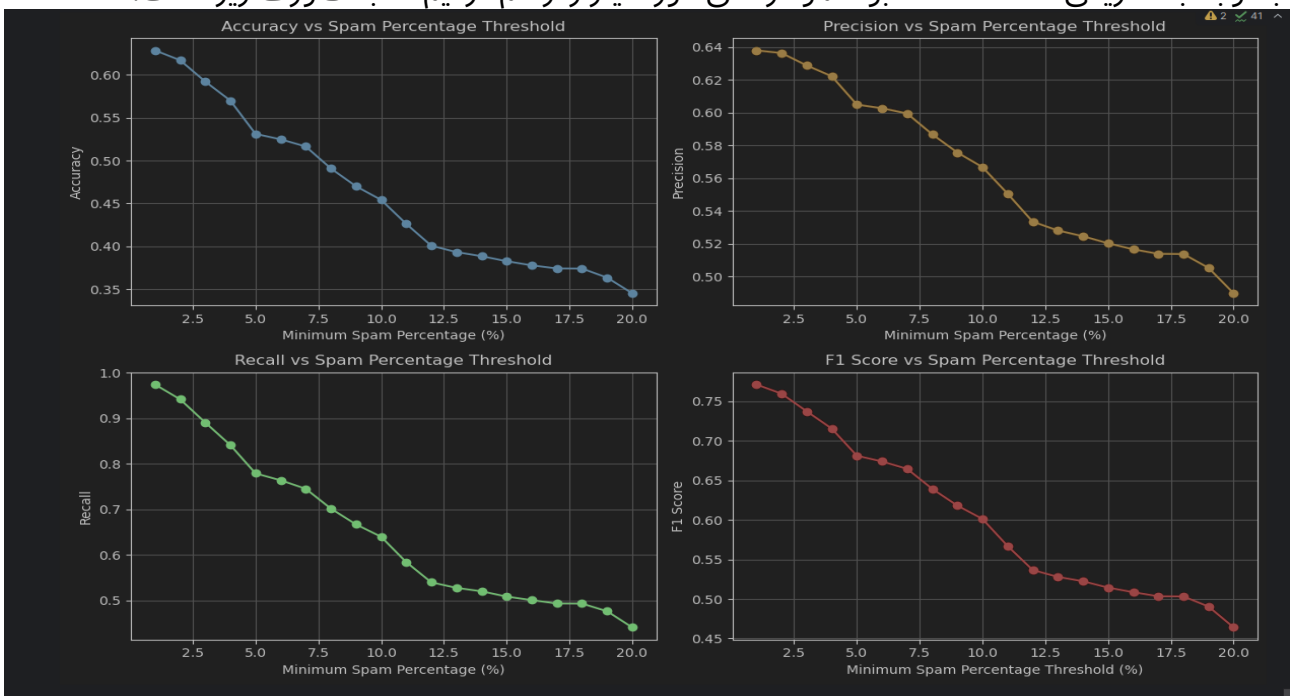



که باز هم میتوان مشاهده کرد درصد بدی به ما میدهد که البته میتوان treshold و اسکور را دستکاری کرد و اون مقدار بهینه شده را بدست آورد شاید بتوان به درصد بهتری برسیم



1.b:

با توجه به متریکی که گفته شده بود نمودار های مورد نیاز را رسم کردیم که به صورت زیر است:




Performance Summary:

| Threshold% | Spam Words | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| 1% | 1152 | 0.628 | 0.638 | 0.974 | 0.771 |
| 2% | 617 | 0.617 | 0.636 | 0.942 | 0.760 |
| 3% | 426 | 0.592 | 0.629 | 0.891 | 0.737 |
| 4% | 352 | 0.570 | 0.622 | 0.841 | 0.715 |
| 5% | 306 | 0.531 | 0.605 | 0.779 | 0.681 |
| 6% | 289 | 0.525 | 0.603 | 0.764 | 0.674 |
| 7% | 237 | 0.516 | 0.600 | 0.745 | 0.665 |
| 8% | 124 | 0.491 | 0.587 | 0.702 | 0.639 |
| 9% | 87 | 0.470 | 0.576 | 0.667 | 0.618 |
| 10% | 66 | 0.454 | 0.567 | 0.640 | 0.601 |
| 11% | 43 | 0.426 | 0.551 | 0.584 | 0.567 |
| 12% | 28 | 0.401 | 0.533 | 0.540 | 0.537 |
| 13% | 25 | 0.393 | 0.528 | 0.528 | 0.528 |
| 14% | 22 | 0.389 | 0.525 | 0.520 | 0.522 |
| 15% | 15 | 0.383 | 0.520 | 0.509 | 0.514 |
| 16% | 14 | 0.378 | 0.517 | 0.501 | 0.509 |
| 17% | 10 | 0.374 | 0.514 | 0.493 | 0.503 |
| 18% | 8 | 0.374 | 0.514 | 0.493 | 0.503 |
| 19% | 6 | 0.364 | 0.505 | 0.476 | 0.490 |
| 20% | 5 | 0.345 | 0.490 | 0.441 | 0.464 |

2.c:

همانطور که یکی از متریک هایی که در کلاس مورد بررسی قرار گرفته است بررسی اثر رخداد چند کلمه درکنار هم بود که یکی از روش‌های این متریک N-gram است. N-gram یک تکنیک پردازش زبان است که ترتیب و ترکیب کلمات را در متن بررسی می‌کند. در این روش، دنباله‌های متوالی از N کلمه (مثلاً ۲کلمه‌ای Bigram یا ۳کلمه‌ای Trigram) استخراج شده و به عنوان ویژگی‌های معنادار برای تحلیل استفاده می‌شوند.