گزارش پایانی فازسوم پروژه

آریانا سیدپیروی 95521198

-۱با توجه به انتخاب شده های نادرست(استقلال انتخاب شده اند ولی در اصل پرسپولیسی هستند.) مثلا سطر سوم فایل متنی واقع در آدرس path\to\git\ClsModel\NaiveBayes\group_samples\fp_esteghlal.txtتوسط classifierبرای استقلال میتوانیم دریابیم که بعضی جمله ها کلمات کلیدی مانند پرسپولیس آمده است اما چون طول جمله زیاد شده و کلمات تعدادشان زیاد شده چند کلمه دیگر آمده اند و باعث شده اند احتمال کلمه پرسپولیس گم شود.پس این جا را می توان از نقطه ضعف های این مدل به شمار آورد.(این مورد در سطر هفتم همین فایل در مورد برانکو نیز صادق است.)با توجه به انتخاب نشده های درست(استقلال انتخاب نشده اند ولی در اصل استقلالی هستند.)همین مورد بالا در سطر بیستم فایل متنی واقع در آدرس path\to\git\ClsModel\NaiveBayes\group_samples\fn_esteghlal.txt

نیز دیده می شود.همچنین در خط ۲۵ این فایل با وجود کلمه ای مثل محسن کریمی که به صورت پیوسته باید در نظر گرفته شود با اینکه در مقایسه با کلمه علی کریمی تعداد کم تری دارید ولی باید متن استقلالی باشد نه پرسپولیسی ولی به دلیل احتمال بالای کریمی این جمله به پرسپولیسی تشخیص داده شده.(البته بنده از فرمول laplace smoothin و در نظرگرفتن کلمه <Unk> فرمول NaiveBayesرا نوشتم که در مخرج1 + |V| داشت.

-۲با توجه به اینکه دیتاست من کاملا متفاوت بود من به تنهایی با تعداد کلمات توانستم نتیجه ی خوبی در این بخش کسب کنم و دقت بالای ۹۰ درصد را به دست بیاورم.به نظر من با توجه به دیتای شخصی معیار های کمی مثل تعداد کلمات آبی‌پوشان یا سرخ‌پوشان یا ... میتوانستند ویژگی های بسیار به درد بخور و حایز اهمیتی باشند که همینطور هم شدویژگی دیگری به درد بخور به ذهنم خطور نکر.(البته خواستم فقط از تعداد کلمات موجود در word cloud ها برای این کار استفاده کنم ولی متوجه شدم کلمات به درد نخور هم ممکن است در بی آن ها یافت شده و حذف شوند و همچنین کلمات پرارزشی که دارای تکرار کمی باشند اما وزن بالایی داشته باشند.

-۳با توجه به دقت بالا من ویژگی های دیگری اضافه نکردم.

-۴به نظر من استفاده از MaxEnt خیلی بهتر بود چون دقت محاسبات را تقریبا به بالای ۹۰٪ رساند و همچنین F1_measureتفاوت بسیار زیادی کرد که در نوع خودش عالی بود ولی با توجه به محاسبات ساده تر من خودم NaiveBayes را ترجیح می‌دهم چرا که سرعت بالایی دارد و تفاوت زیادی با MaxEnt ندارد ولی شاید در دیتاهای بیشتر با چالش های بیشتر مثلا جاهایی که ترتیب کلمات خیلی مهم باشند یا جمله های طولانی تری باشند استفاده از MaxEnt دقت را بسیار بالا

ببرد.البته من در فیچرهایم کلمه <Unk> را هم در نظر گرفتم چون کلمه ای است که در train data دیده نشده است.