باسمه تعالی

گزارش تمرین شماره ۴ درس پردازش زبان طبیعی

استاد درس: جناب آقای دکتر باباعلی

نام دانشجو : ایمان کیانیان

شماره دانشجویی: ۶۱۰۳۰۰۲۰۳

۱- مقدمه

در این تمرین قصد داریم عمل دسته بندی را روی dataset پرسیکا انجام دهیم. دیتاست پرسیکا، یک دیتاست شامل ۱۱۰۰۰ متن خبری است که از سایت ایسنا جمع آوری شده است. این متون در دسته های کلی مختلف (۱۱ دسته) جمع آوری شده اند که دسته های مربوطه به صورت زیر هستند:

- آموزشی
- اجتماعی
- اقتصادی
- بهداشتی
- تاریخی
- سیاسی
- علمی
- فرهنگی
- فقه و حقوق
- مذهبی
- ورزشی

از هر دسته ۱۰۰۰ خبر وجود دارد. ابتدا روی این اخبار پیش پردازش انجام میدهیم. سپس داده های تست و آموزشی را جدا میکنیم و با داده های آموزشی و تست را با استفاده از مفهوم tf-idf به دو ماتریس با word-document تعداد سطر های یکسان تبدیل میکنیم. سپس با استفاده از تجزیه SVD، حجم این ماتریس sparse و بزرگ را کم میکنیم تا بتوانیم به راحتی عمل دسته بندی را روی این دادگان انجام دهیم.

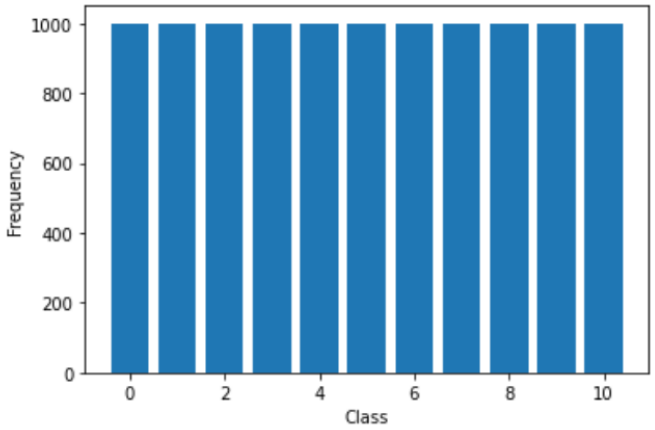
۲- پیش پردازش داده ها

برای پیش پردازش داده ها تکنیک های زیادی باهم و بدون هم استفاده کردیم. از جمله تست های مختلف میتوان به اعمال stemmer ، normalizer ، حذف کردن stop word های فارسی از متن و ... اشاره کرد. در تست های مختلف دقت های مختلفی را نیز گرفتیم که در اینجا به نتایج خوب اشاره میکنیم.

در ابتدا دیتا را لود کردیم، سپس ضایعات دیتا را حذف کردیم و دیتا را به یک dataframe در pandas تبدیل کردیم. سپس قسمت label که به آن نیاز داریم را با استفاده از labelencoder تبدیل به عدد های صحیح کردیم یعنی مثلا اخبار متناظر با موضوع آموزشی کلاس صفر هستند، اخبار داده های اجتماعی کلاس ۱ و ... .

```
The number corresponding to the آموزشي is:        0
The number corresponding to the اجتماعي is:       1
The number corresponding to the اقتصادي is:       2
The number corresponding to the بهداشتي is:       3
The number corresponding to the تاريخي is:        4
The number corresponding to the سياسي is:         5
The number corresponding to the علمي is:          6
The number corresponding to the فرهنگي is:        7
The number corresponding to the فقه و حقوق is:    8
The number corresponding to the مذهبي is:         9
The number corresponding to the ورزشي is:         10
```

سپس تعداد هر یک از این کلاس ها را میشماریم و به صورت نمودار رسم میکنیم:



سپس دادگان stop word را که از اینترنت دانلود کردیم(در فایل stop.txt ارسال شده است) لود کردیم. پیش پردازش ساده از از جمله نرمال کردن و اعمال stem روی آن ها را انجام دادیم. این مجموعه stop word شامل ۳۸۹ کلمه است که از جملات آموزشی و تست حذف میکنیم. سپس نماد های زبان را به صورت جدا به صورت زیر تعریف کردیم و از متن آموزشی و تست حذف کردیم. نماد هایی که حذف شد به صورت زیر است:

```
['!','»','،','؟','؛','"','#','(',')','*',',','-','.','/',':','[',']','«','…']
```

با استفاده از word tokenizer پکیج hazm، کلمات را برای هر خبر جدا کردیم، کمی نرمالش کردیم (رعایت نیم فاصله ها و یکسان سازی فرم نوشتار و ...) و سپس دوباره تبدیل به خبر کردیم. سپس وارد مراحل بعد که تشکیل ماتریس tf-idf و استفاده از svd است شدیم.

۳- تشکیل ماتریس TF-IDF و استفاده از SVD

بعد از پیش پردازش داده های آموزشی و تست ، حذف حروف اضافه و نماد ها و ...، وارد بخش تبدیل متون خبری به ماتریس TF-IDF شدیم. این ماتریس را با استفاده از پکیج sklearn انجام دادیم. تابع TfidfVectorizer مورد استفاده ما است. داده های تست و آموزشی را با استفاده از این تابع به یک ماتریس word-document تبدیل کردیم که sparse است. سپس این ماتریس را با کمک یک SVD با تعداد مولفه ۵۰۰ تبدیل کردیم تا فشرده سازی ماتریس به خوبی انجام شود. ماتریس اولیه تعداد feature بالغ بر ۶۲۸۶۵ داشت که با کمک SVD تبدیل به ۵۰۰ شد. حال داده های ما آماده برای انجام عمل classification هستند. البته لازم به ذکر است که بر روی تعداد پارامتر های کمتر و بیشتر SVD (بزرگتر و یا کوچکتر از ۵۰۰) نیز تست انجام دادیم. هر چه تعداد مولفه های SVD بیشتر باشد پیچیدگی محاسباتی بیشتر و دقت بیشتری خواهیم داشت و هر چه تعداد مولفه ها کمتر باشد، پیچیدگی محاسباتی کمتر ولی دقت گرفته شده از classifier ها کمتر خواهد بود بنابراین باید یک tradeoff را هنگام انتخاب تعداد مولفه های SVD رعایت کنیم.

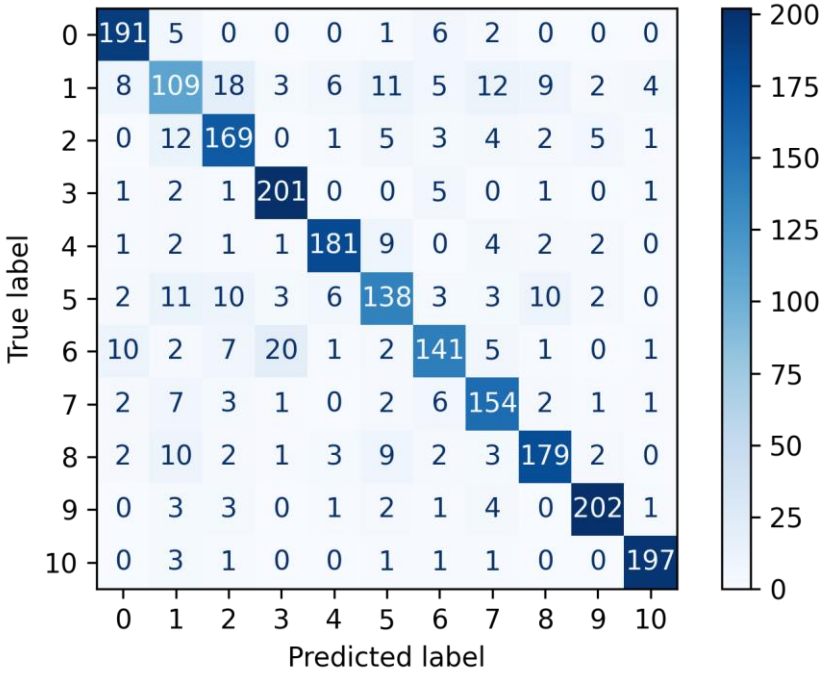
۴- اعمال classifier های مختلف

- **اعمال SVM با کرنل rbf** :

یک SVM با کرنل rbf را آموزش دادیم و دقت آن را روی داده های تست و آموزشی گرفتیم که روی داده های آموزشی دقت برابر ۰,۹۳۷۶۰۶۵۴۶۱۹۸۴۳۱۶ و دقت بر روی داده های تست برابر ۰,۸۴۶۳۶۳۶۳۶۳۶۳۶۳ است.

```
Accuracy on Train Data = 93.76065461984317%
Accuracy on Test Data = 84.63636363636363%
```

Confusion matrix برای داده های تست:



که لیبل های ذکر شده به صورت عددی همان لیبل هایی هستند که در ابتدای گزارش توافق کردیم مثلا اخبار آموزشی عدد ۰ ، اخبار اجتماعی ۱ و ... است. با تحلیل نتایج در این واضح است که سخت ترین موضوع برای دسته بندی اخبار اجتماعی است. مقادیر دقت ها با جزئیات بیشتر را میتوانید مشاهده کنید:

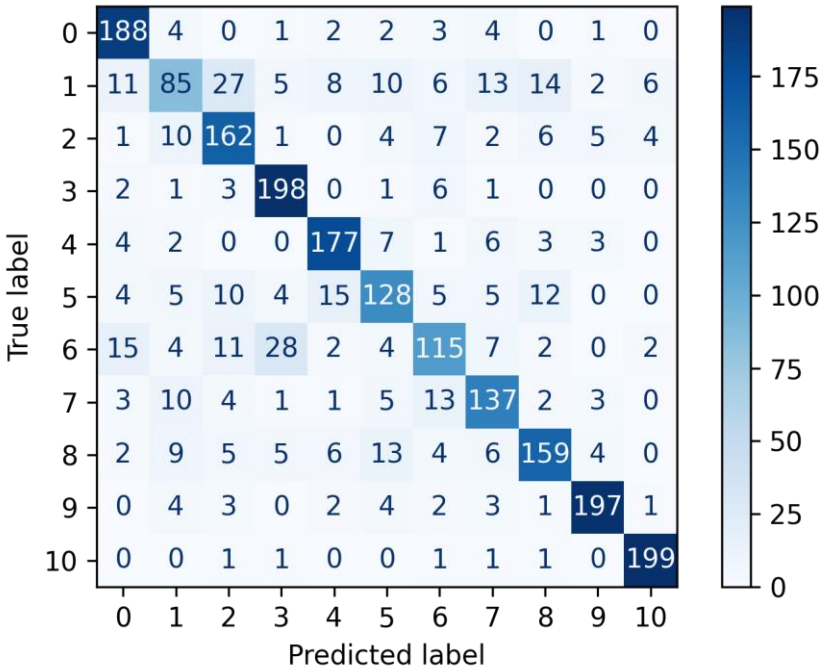
	precision	recall	f1-score	support
0	0.88	0.93	0.91	205
1	0.66	0.58	0.62	187
2	0.79	0.84	0.81	202
3	0.87	0.95	0.91	212
4	0.91	0.89	0.90	203
5	0.77	0.73	0.75	188
6	0.82	0.74	0.78	190
7	0.80	0.86	0.83	179
8	0.87	0.84	0.85	213
9	0.94	0.93	0.93	217
10	0.96	0.97	0.96	204
accuracy			0.85	2200
macro avg	0.84	0.84	0.84	2200
weighted avg	0.84	0.85	0.84	2200

همانطور که مشخص است برای کلاس دوم که همان اخبار اجتماعی است، مقدار recall تقریبا پایینی داریم که یعنی ۵۸ درصد از اخبار اجتماعی ، به درستی اجتماعی تشخیص داده شده اند که مشخصا این مقدار خوب نیست. با استفاده از confusion matrix، میتوانیم دریابیم که در کلاس ۱، بیشتر با کلاس های ۳ ، ۵ و ۷ تداخل دارد. احتمالا در کلاس اجتماعی، اخباری داریم که بسیار شبیه به موضوعات دیگر هستند یا کلمات کلیدی مشترک دارند.

همچنین لازم به ذکر است SVM با پارامتر ها و کرنل های مختلفی تست شد اما در اینجا بهترین نتیجه گرفته شده از SVM را آورده ایم.

- اعمال **random forest** :

بعد از اعمال random forest با n_estimator = 100 ، دقت بر روی داده آموزشی ۹۹,۹۵۴۵۴۰۲۸۸۶۶۹۱۶٪ و دقت روی داده تست برابر ۷۹,۳۱۸۱۸۱۸۱۸۱۸۳٪ بود که درگیر overfit هستیم. Confusion matrix بر روی داده تست به صورت زیر است:



همچنین دقت های گرفته شده به صورت جزئی در این قسمت به صورت زیر است:

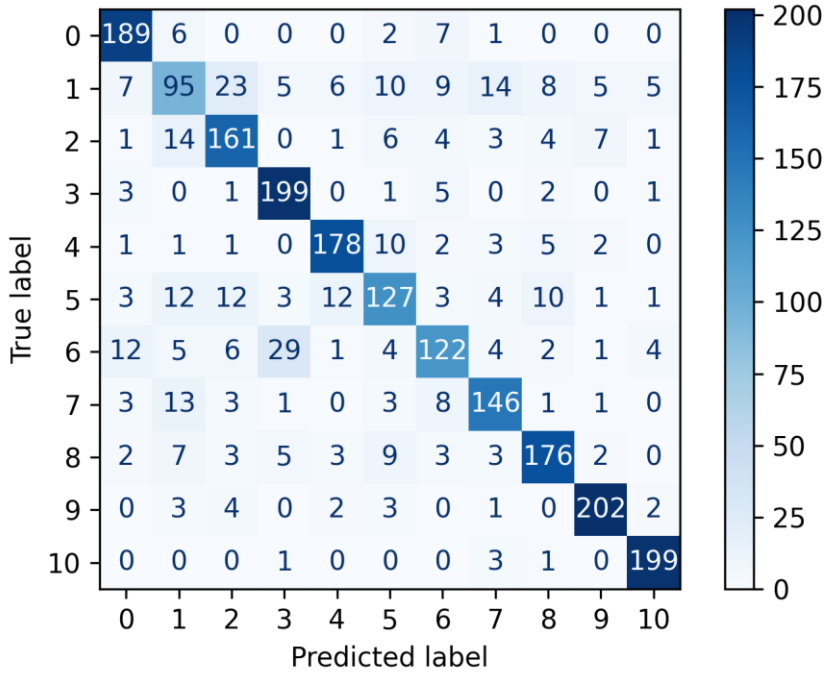
	precision	recall	f1-score	support
0	0.82	0.92	0.86	205
1	0.63	0.45	0.53	187
2	0.72	0.80	0.76	202
3	0.81	0.93	0.87	212
4	0.83	0.87	0.85	203
5	0.72	0.68	0.70	188
6	0.71	0.61	0.65	190
7	0.74	0.77	0.75	179
8	0.80	0.75	0.77	213
9	0.92	0.91	0.91	217
10	0.94	0.98	0.96	204
accuracy			0.79	2200
macro avg	0.78	0.79	0.78	2200
weighted avg	0.79	0.79	0.79	2200

همانطور که مشاهده می شود باز هم در recall برای کلاس "اجتماعی" مشکل داریم.

- **اعمال XGBoost** :

بعد از اعمال XGBoost دقت برای داده های آموزشی برابر ٪۹۹,۹۵۴۵۴۰۲۸۸۶۶۹۱۶ و برای داده های تست به صورت ٪۸۱,۵۴۵۴۵۴۵۴۵۴۵۴۵۵ بود که مشخصا باز هم درگیر overfit شده ایم.

ماتریس آشفتگی برای داده های تست به صورت زیر است:

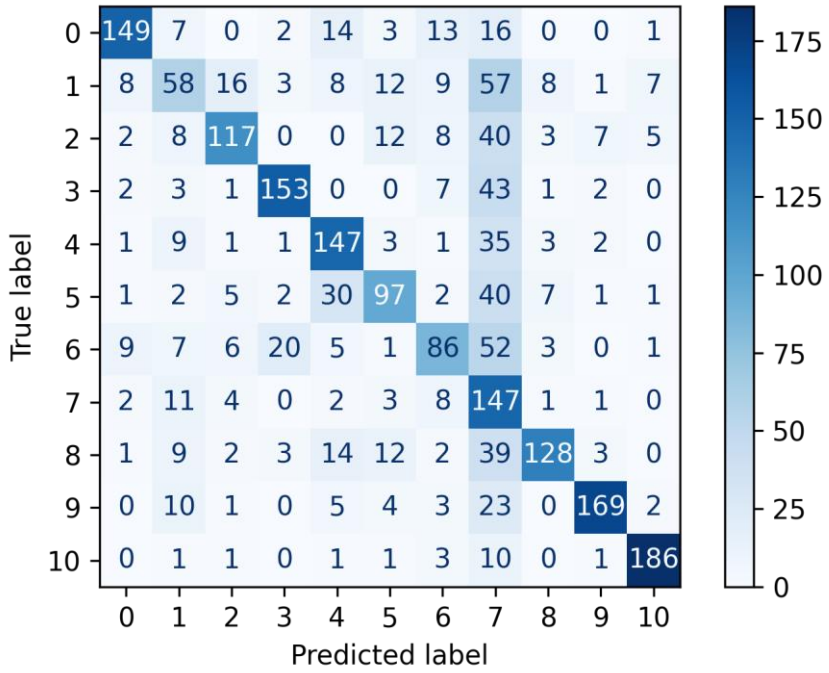


باز هم طبق انتظار و حدس، در کلاس ۱ که همان اخبار اجتماعی است مشکل داریم. دقت های گرفته شده در تصویر زیر نیز گواه این قضیه است:

	precision	recall	f1-score	support
0	0.86	0.92	0.89	205
1	0.61	0.51	0.55	187
2	0.75	0.80	0.77	202
3	0.82	0.94	0.87	212
4	0.88	0.88	0.88	203
5	0.73	0.68	0.70	188
6	0.75	0.64	0.69	190
7	0.80	0.82	0.81	179
8	0.84	0.83	0.83	213
9	0.91	0.93	0.92	217
10	0.93	0.98	0.95	204
accuracy			0.82	2200
macro avg	0.81	0.81	0.81	2200
weighted avg	0.81	0.82	0.81	2200

- اعمال **Gaussian Naïve bayes** :

بعد از اعمال Naïve bayes دقت ۷۰٫۸۳۷۵۹۵۱۸۱۲۷۰۶٪ و ۶۵٫۳۱۸۱۸۱۸۱۸۱۸۱٪ را به ترتیب برای داده های آموزشی و تست گرفتیم. از نتیجه میتوان دریافت که این مدل underfit شده است. در واقع علت آن این است که مدل توان مدلسازی بالایی برای این داده ها ندارد. ماتریس آشفتگی برای داده های تست در این حالت به صورت زیر است:



همانگونه که مشاهده میشود مدل سعی کرده تمامی کلاس هارا، کلاس ۷ پیش بینی کند و بین آنها تفاوت چندانی قائل نمیشود. همچنین مشکل recall در کلاس ۱ یا اخبار اجتماعی همچنان وجود دارد.

	precision	recall	f1-score	support
0	0.85	0.73	0.78	205
1	0.46	0.31	0.37	187
2	0.76	0.58	0.66	202
3	0.83	0.72	0.77	212
4	0.65	0.72	0.69	203
5	0.66	0.52	0.58	188
6	0.61	0.45	0.52	190
7	0.29	0.82	0.43	179
8	0.83	0.60	0.70	213
9	0.90	0.78	0.84	217
10	0.92	0.91	0.91	204
accuracy			0.65	2200
macro avg	0.71	0.65	0.66	2200
weighted avg	0.72	0.65	0.67	2200

لازم به ذکر است که اعمال Multinomial Naïve bayes ممکن نبود زیرا در تجزیه SVD داده های منفی نیز داریم که قابل قبول نیست. داده ها را در بازه ۰ و ۱ نرمال کردیم ولی باز هم از MN NB دقت بالایی نگرفتیم.

۵- بهترین دقت دریافت شده

همانطور که ذکر شد در حالتی که از SVM با کرنل rbf استفاده کردیم بیشترین دقت را از مدل برای داده تست که تقریبا برابر **۸۵ درصد** بود گرفتیم. این دقت با توجه به داده های ما و اینکه با استفاده از SVD تجزیه انجام میدهیم قابل قبول است.

۶- جمع بندی

در این تمرین، بررسی های لازم جهت دریافت دقت های بیشتر انجام شد. سعی کردیم تمام حالات مختلف را برای پیش پردازش داده و انواع پارامتر های ممکن را برای انواع دسته بند ها ، tfidf و svm را تست کنیم. نتایج و آزمایش هایی که در این فایل گزارش آورده ایم نتایج حاصل از آزمایش های متعدد است اما به دلیل تعداد بالای آزمایشات آوردن آن در فایل گزارش ممکن نبود. بهترین دقت را در SVM گرفتیم که میتوانید جزئیات بیشتر را در قسمت های قبل مشاهده کنید.