به نام خدا

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده برق و کامپیوتر

درس پردازش زبان های طبیعی

تمرین ۲

سجاد پاکدامن ساوجی
۸۱۰۱۹۵۵۱۷

فروردین ۱۴۰۰

فهرست مطالب

۱ - پیش پردازش

در این قسمت با توجه به خواسته سوال منشن ها و ایموجی ها حذف شدند و هشتگ ها باز شدند. همچنین برای بهبود عملکرد مدل contraction ها نیز اصلاح شدند. برای مثال کلمه کوتاه شده don't به do not تبدیل شده است. همچنین عملیات stemming و lemmatization نیز روی کلمات اعمال شده است.

پس از پیش‌پردازش متون با استفاده از word cloud کلمات پرتکرار در تصویر ۱ آورده شده است.

شکل ۱. کلمات پرتکرار در مجموعه داده

از آن جایی که مجموعه داده متوازن نیست، لازم است با روشی توازن را در این مجموعه داده برقرار کنیم. برای این کار روش های متفاوتی وجود دارد که در ادامه چندی از آن ها توضیح داده میشوند.

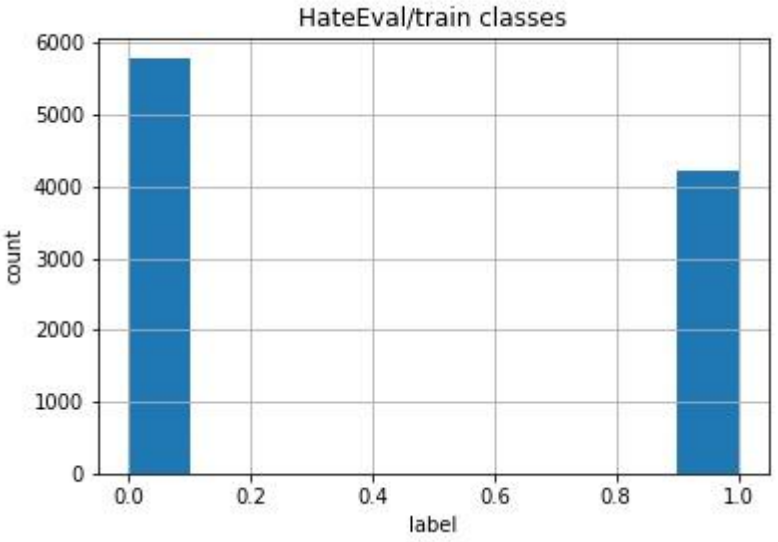
۱. **استفاده از داده های کمکی:** این روش موثر ترین روش و در عین حال سخت ترین روش است. چرا که اگر تعداد داده بیشتری موجود باشد، مجموعه داده میتواند نماینده بهتری از کل جامعه آماری باشد. در تمرین پیش داده های کمکی در اختیار ما قرار داده شده است و میتوان از آنها استفاده کرد.

۲. **تولید داده های جدید:** در این روش از یک روش جنریتیو برای تولید داده های جدید استفاده میشود. برای مثال در NLP میتوان از ترجمه ماشینی برای تولید داده های جدید استفاده کرد.

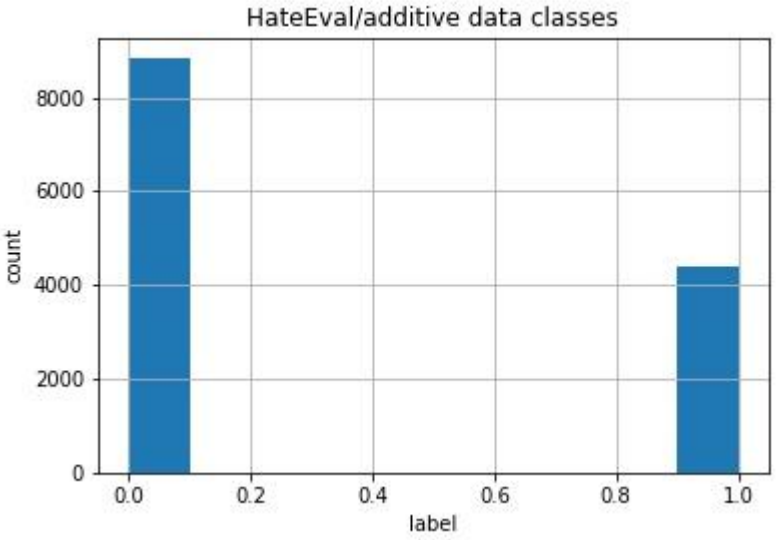
۳. **محدود کردن مجموعه داد:** در این روش، داده هایی از کلاس پر جمعیت تر حذف می شود تا تعداد داده های هر دو کلاس برابر شوند.

۴. **روش oversampling و augmentation:** در این روش ها از داده های موجود استفاده می شود تا داده هایی بسیار نزدیک به داده های اصلی تولید شود. برای مثال حذف قسمتی از متن و یا تغییر ترتیب جملات میتواند یکی از روش های augmentation باشد.

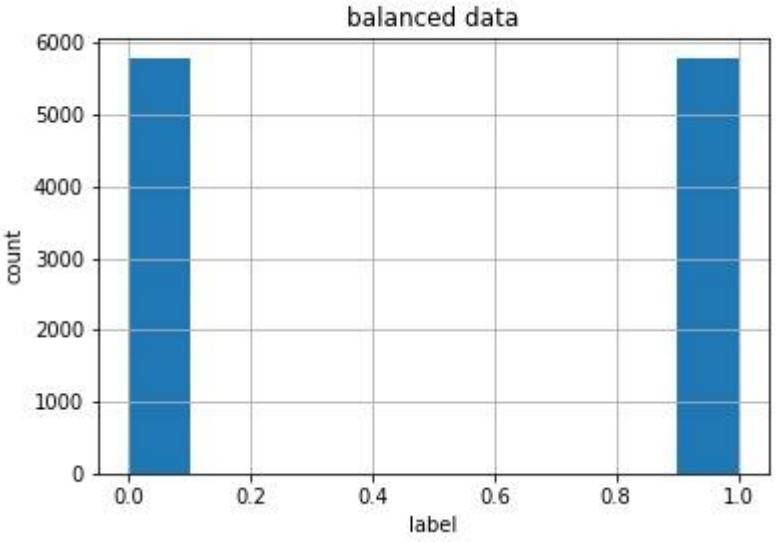
در این تمرین از روش ۱ برای متوازن کردن مجموعه داده استفاده شده است. در شکل ۲، ۳ و ۴ هیستوگرام داده ها پیش و پس از متوازن سازی آمده است.



شکل۲. هیستوگرام برچسب داده های آموزش



شکل۳. هیستوگرام برچسب داده‌های اضافی



شکل۴. هیستوگرام داده های متوازن شده

۲ - استخراج ویژگی

.

در این قسمت تلاش شده است که تنها از frequency کلمات مختلف استفاده شود. به این منظور تعدد تکرار کلمات را در دو کلاس در داده های آموزش ابتدا بررسی کرده و با استفاده از معیار Information Gain آن ویژگی هایی را انتخاب میکنیم که IG بیشتری داشته باشند. این روش در انتخاب ویژگی جدا کننده در decision tree نیز استفاده میشود، با این تفاوت که در DT در هر گره از درخت این معیار بر قسمتی از مجموعه داده اعمال میشود.

در نهایت که کلمات مناسب انتخاب شده است، برای استخراج ویژگی تعداد تکرار این کلمات خاص را در متن بدست می آوریم و مقدار log این بردار را به عنوان بردار ویژگی انتخاب میکنیم.

$$E = -\sum_{i}^{C} p_i \log_2 p_i$$

جدای این روش ، ویژگی های دیگری نیز انتخاب شد، اما این روش عملکرد بهتری داشت. تاثیر این ویژگی ها نیز بر روی طبقه بندی سنجیده شده است. در حالتی که بر اساس IG ویژگی ها انتخاب شده اند، آن ویژگی که IG بیشتری داشته باشد در طبقه بندی نیز اهمیت بیشتری خواهد داشت که از تعریف IG نیز همین پیشبینی میشد.

۳ - آموزش رده بندی

در این قسمت دو طبقه بند naive bayes و logistic regression بر روی داده ها آموزش داده شده است. برای بدست آوردن معیار بهتری از عملکرد مدل از fold-5 استفاده شده است. در جدول شماره ۱ عملکرد مدل گزارش شده است.

model	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean	std
Guassian Naive Bayes	0.57642	0.48359	0.57599	0.67918	0.65371	0.5937	0.068
Logistic Regression	0.56174	0.55094	0.59110	0.66925	0.67055	0.608	0.051

جدول ۱. عملکرد مدل‌ها روی داده های آموزش

علت استفاده از k-fold آن است که با استفاده از آن (و دیگر روش های cross validation) می توان مشکل overfitting را تشخیص داد. چون که در هر گام از cross validation مدل روی قسمت خاصی آموزش داده میشود، اگر دقت بر روی داده های تست پایین باشد و روی قسمت داده های آموزش بالا باشد، به معنی overfitting است.

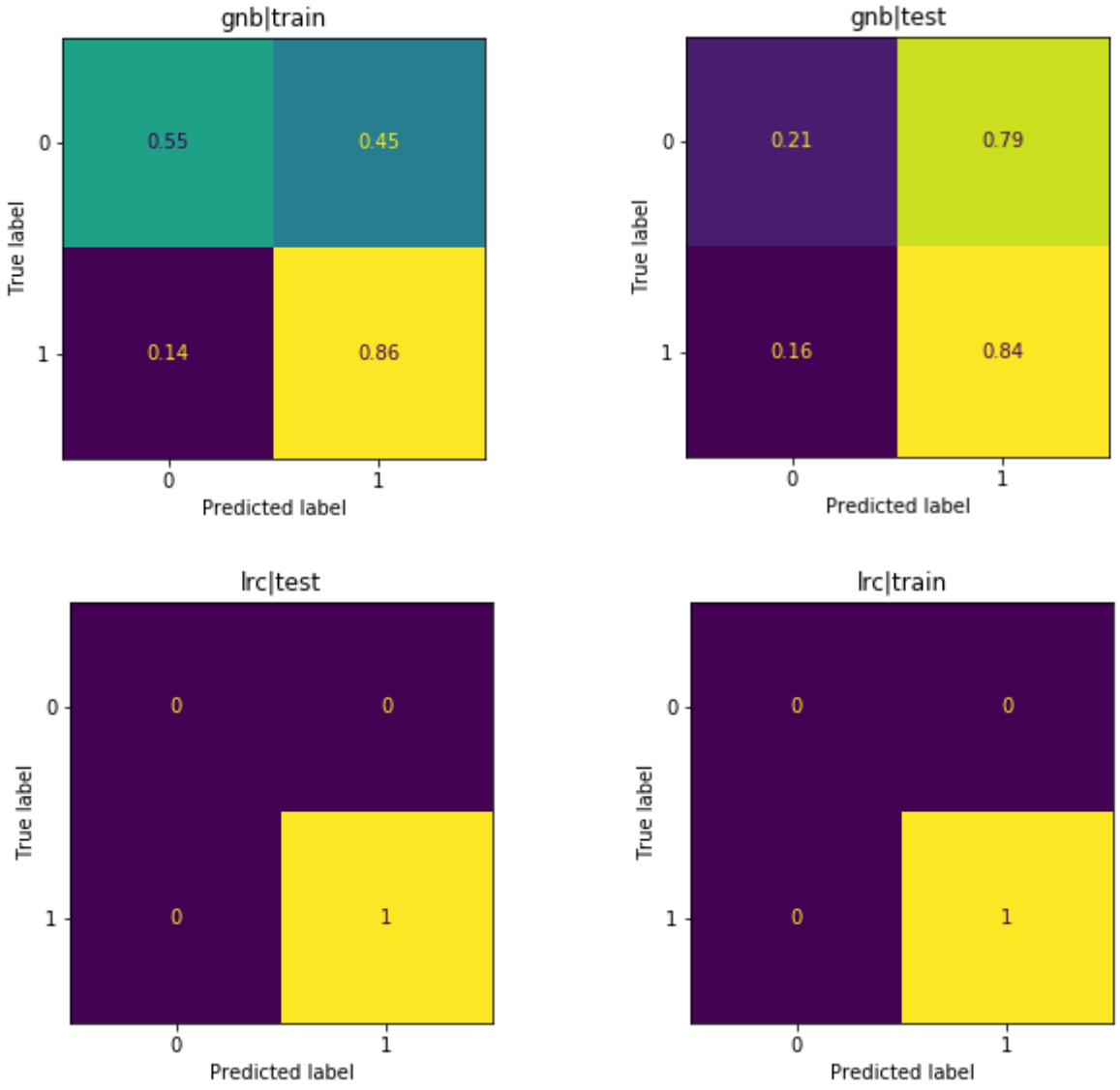
۴ - ارزیابی مدل

مطابق خواسته های سوال، معیار های زیر در جدول ۲ نشان دهنده عملکرد مدل بر روی داده های آزمایش و آموزش هستند.

model	acc	perc	recall	f1-micro	f1-macro
gnb - train	0.7064766	0.6576552	0.8613126	0.6992668	0.7064766
lrc - train	0.8126079	0.8193366	0.8020725	0.8125871	0.8126079
gnb - test	0.4773333	0.4366255	0.8420634	0.4481399	0.4773333
lrc - test	0.5306666	0.4675723	0.8468253	0.5148305	0.5306666

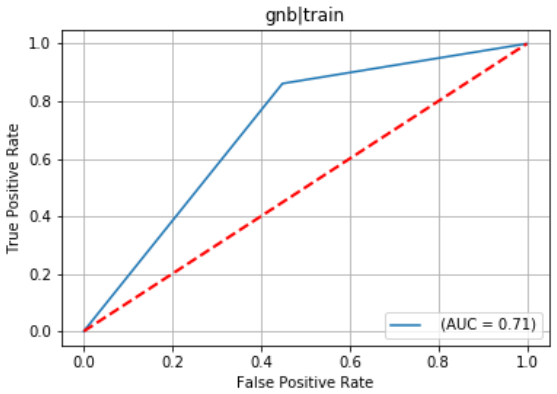
جدول ۲. عملکرد طبقه بندی بر روی داده های آموزش و آزمایش

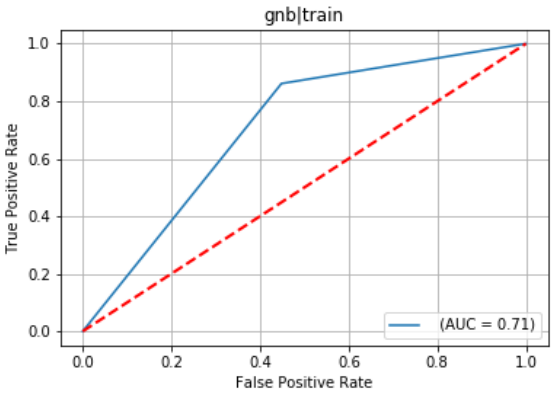
با توجه به جدول ۲، طبقه بند naive bayes در این سوال عملکرد و تعمیم دهی بهتری داشته است. همچنین ماتریس های پراکندگی این طبقه بندها در تصویر ۵ و نمودار های ROC برای این طبقه بند ها در تصویر ۶ آمده است.



شکل ۵. ماتریس های پراکندگی برای طبقه بند های مختلف

در تصویر ۶ نمودار های ROC برای دو مدل آورده شده است.





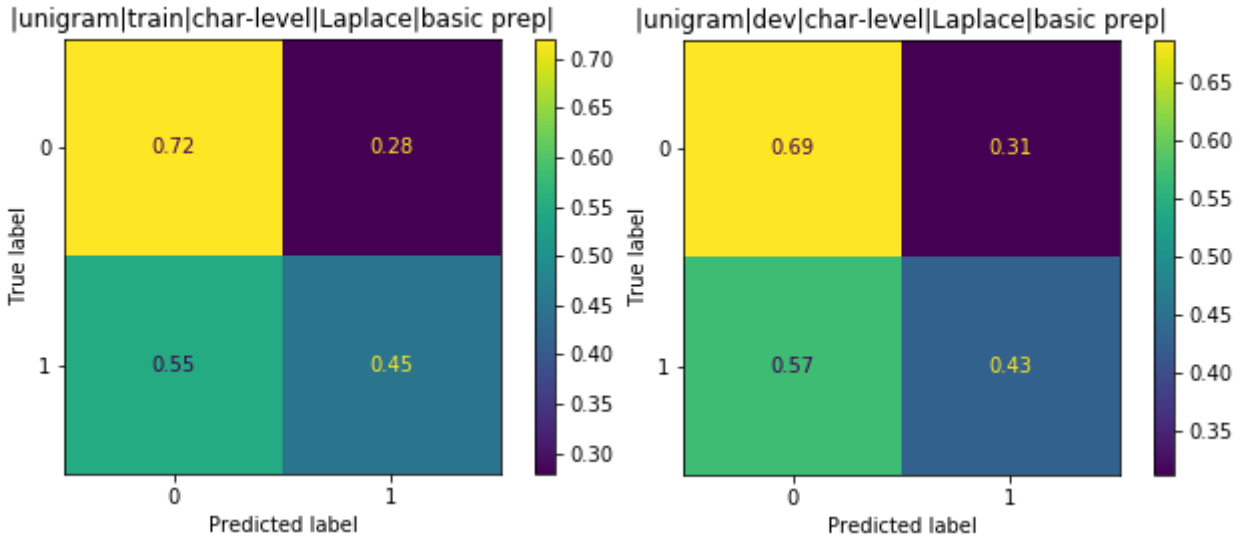
شکل ۶. نمودار های roc برای طبقه بند ها

همچنین برای مقایسه عملکرد مدل ها با مدل های تمرین قبلی در جدول ۳ معیار های ارزیابی مدل های قبلی آورده شده است.

model	F1-macro	F1-micro	acc	prec	recall
unigram|word	0.8951963615	0.8954379911	0.8954379911	0.8607904861	0.9434540923
bigram|word	0.9832245790	0.9832279087	0.9832279087	0.9699850857	0.9973164654
unigram|char	0.5776370987	0.5853939045	0.5853939045	0.6171443597	0.4498754073
bigram|char	0.6882800772	0.6884224650	0.6884224650	0.6806985294	0.7097949012

جدول ۳. عملکرد مدل روی داده های آموزش در تمرین ۱

در ادامه ماتریس های پراکندگی مدل های تمرین ۱ در شکل ۷ آورده شده است.



شکل ۷. ماتریس های پراکندگی در تمرین ۱

نتیجه گیری و مقایسه مدل ها:

همانطور که از داده های جدول ۱،۲ و ۳ مشخص است، عملکرد مدل هایی که در تمرین ۱ استفاده کرده ایم، بهتر بوده است. این مدل ها در داده های آموزش و آزمایش بهتره عمل کرده اند که نشان دهنده قدرت تعمیم دهی بالاتر و اطمینان بهتری است.

علت بهتر بودن مدل های پیشین را میتوان استخراج دستی ویژگی ها دانست. در حالی که در مدل های ngram نیازی به استخراج ویژگی نبود، اما در این تمرین باید ویژگی ها طراحی می شوند، لذا انتخاب ویژگی اهمیت ویژه ای پیدا میکند.

احتمال دارد که با استخراج بهتر ویژگی به دقت های بالاتری رسید، اما در این تمرین با این که ویژگی های متعددی آزمایش شد، مدل های آموزش دیده شده نتوانستند دقت های بالاتری بدست آورند.