گزارش کار تمرین دوم-طبقه بندی متن

محمد لشکری ۴۰۰۱۱۲۰۸۷

۲۲ فروردین ۱۴۰۱

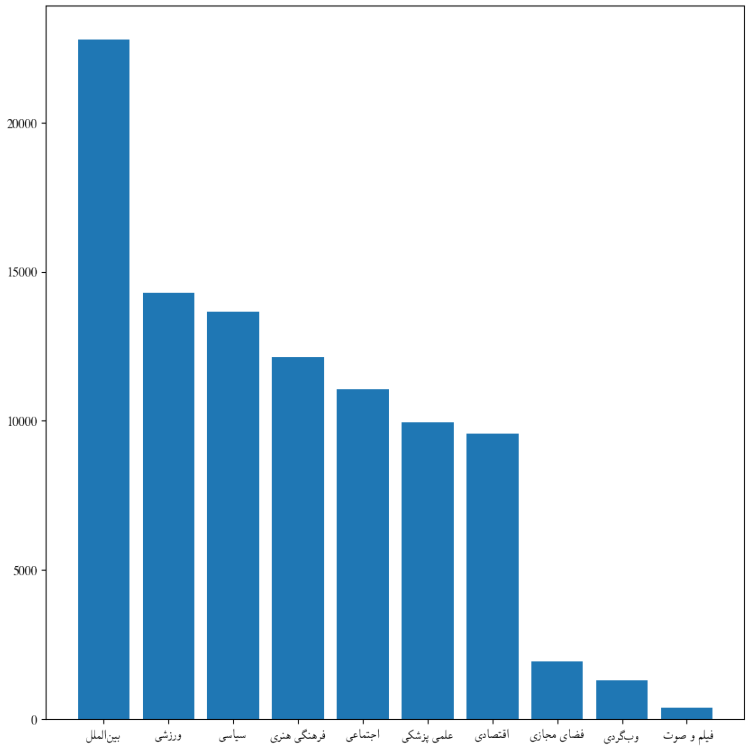
۱ پیش‌پردازش دادگان

در تابع clean_text() تمامی کاراکترها به‌جز نقطه، علامت سوال، حروف فارسی، اعداد و فاصله حذف شدند. همچنین اعداد با کاراکتر N جایگزین شدند و کاراکتر \xa0 نیز از دادگان حذف شده است.

تابع count_words() تعداد تکرار هر توکن منحصر به فرد را در دیکشنری frequencies ذخیره می‌کند و بعد از مرتب‌سازی، ۲۰۰ توکن پرتکرار را در فایل frequent.txt ذخیره می‌کند. سپس تعداد توکن‌ها چاپ می‌شود. تعداد توکن‌های منحصر به فرد و تعداد کل توکن‌های مجموعه آموزشی به ترتیب ۴۳۷,۱۸۳ و ۳۲,۰۵۸,۵۱۱ و برای مجموعه آزمایشی به ترتیب ۱۴۳,۳۶۵ و ۵,۷۹۴,۵۸۶ است. لازم به ذکر است نمونه هایی که برچسب با نام category داشتند که تعداد آنها ۸ بوده از دادگان حذف شده‌اند. سه نمونه از نتاظرهای انجام شده در جدول ۱ قابل مشاهده است.

کلمه	اندیس
گفت	۲۰
هم	۳۰
مردم	۵۰

جدول ۱: نمونه‌های تناظرهای انجام‌شده



شکل ۱: توزیع دادگان آموزشی

۲ طبقه‌بندی کننده بیز ساده

در تابع count_word _per_class() یک دیکشنری به نام count_per_class_dict تعداد تکرار هر کلمه در هر کلاس را نگهداری می‌کند. دیکشنری count_allwords_per_class تعداد تکرار همه کلمات موجود در هر کلاس را نگهداری می‌کند. لگاریتم احتمال‌های prior نیز در این تابع محاسبه و در لیست log_prior_list ذخیره می‌شود و در تابع calculate_log_prior() فقط مقدار آن بازگردانده می‌شود تا مدل کاراتر [1] باشد.

۱.۲ ارزیابی

نتایج حاصل شده برای مجموعه‌های آموزشی و آزمایشی در دو جدول زیر قابل مشاهده است:

از آنجا که توزیع داداگان متوازن نیست، بهترین معیار ارزیابی میانگین ماکروی F1 است. F1 دادگان آمورشی به F1 نزدیک است که نشان می‌دهد مدل بیش از حد آمورش ندیده است. مقادیر recall برای سه

[1] More efficient

	precision	recall	f1-score	support
0	0.93	0.95	0.94	4105
1	0.98	0.97	0.97	2506
2	0.82	0.82	0.82	2473
3	0.86	0.91	0.89	2191
4	0.81	0.79	0.80	2002
5	0.85	0.90	0.88	1821
6	0.87	0.83	0.85	1745
7	0.65	0.71	0.68	339
8	0.67	0.33	0.45	239
9	1.00	0.01	0.03	73
accuracy			0.88	17494
macro avg	0.84	0.72	0.73	17494
weighted avg	0.87	0.88	0.87	17494

(آ) نتایج دادگان آزمایشی

	precision	recall	f1-score	support
0	0.94	0.96	0.95	22767
1	0.98	0.98	0.98	14274
2	0.84	0.85	0.84	13647
3	0.90	0.93	0.91	12136
4	0.85	0.82	0.84	11069
5	0.88	0.91	0.90	9938
6	0.89	0.87	0.88	9583
7	0.69	0.74	0.71	1936
8	0.83	0.48	0.60	1297
9	0.75	0.03	0.06	375
accuracy			0.90	97022
macro avg	0.85	0.76	0.77	97022
weighted avg	0.90	0.90	0.89	97022

(ب) نتایج دادگان آموزشی

کلاس با کمترین فراوانی از سایر کلاس‌ها کمتر است که نشان می‌دهد مدل به سمت کلاس‌های با فراوانی بالاتر bias شده است.