علی اکبر احراری

مستندات پروژه طبقه‌بندی خبرهای فارسی

در این پروژه به آموزش یک مدل برای طبقه‌بندی خبرهای فارسی پرداختیم. دیتاست گزینه اول برای انجام این کار انتخاب شد.

دانلود از این لینک :

dataset/data-news-s/parsaabdolmaleki/persionthttps://www.kaggle.com/dataset

روش پیش‌پردازش

داده‌ها از فایل CSV بارگذاری و ستون‌های غیرضروری حذف شدند. عنوان و توضیحات در ستون `text` ترکیب و مقادیر خالی حذف شدند. متن‌ها باhazm نرمال‌سازی، توکن‌سازی و لماتایز شدند؛ کلمات توقف حذف و نویزهایی مانند URL و کاراکترهای غیرفارسی با re پاک‌سازی شدند. متن‌ها با TfidfVectorizer حداکثر 5000ویژگی، تک‌واژه و دوواژه، min_df=2به بردار تبدیل شدند و عدم تعادل کلاس‌ها با SMOTE رفع شد.

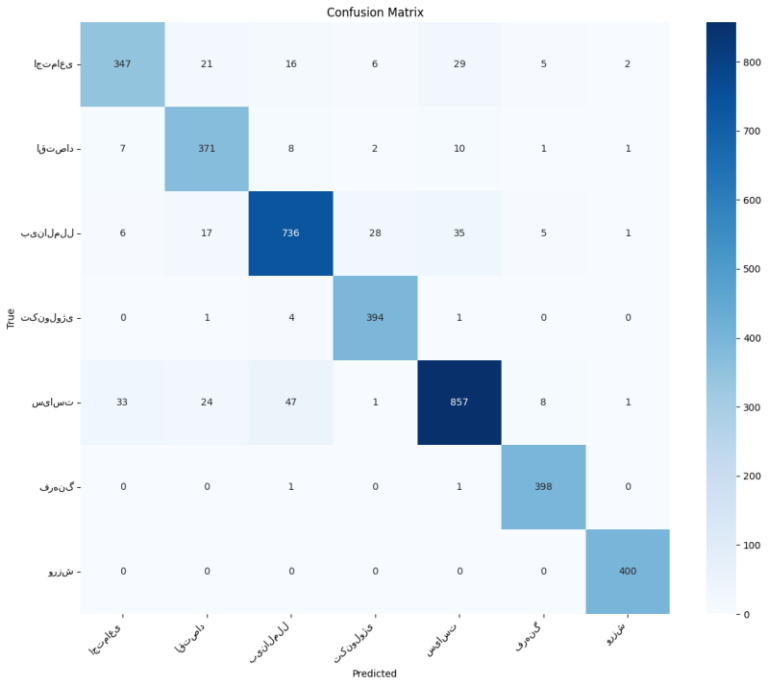
انتخاب مدل و تنظیمات

مدل LinearSVC به دلیل کارایی در داده‌های متنی انتخاب شد. تنظیمات شامل `class_weight='balanced'` برای رفع عدم تعادل، max_iter=10000 و C=1.0 بود. با GridSearchCV و جستجوی C=[0.8, 1, 1.2] ، بهترین مقدار C=1.2 با دقت اعتبارسنجی 91 به دست آمد. داده‌ها با نسبت 80-20 و stratify تقسیم شدند.

نتایج عددی و ماتریس سردرگمی

نتایج کلی:

accuracy	0.915817	0.915817	0.915817	0.915817
macro avg	0.916997	0.927648	0.921596	3825.000000
weighted avg	0.915811	0.915817	0.915210	3825.000000



چالش‌ها و پیشنهادات

مشکل لیبل‌ها: اولین چالش به وجود آمده در این کد مربوط به بود. مشاهده شد که تعداد لیبل ها بسیار زیاد بوده و فرمت و نامشان به اشتباهی تعریف شده بود. پس نیاز به یک پاکسازی عمیق و نسبت دادن درست این لیبل‌ها به داده متناظرشان بود. برای این کار، با توجه به لیبل های دیتاست دیگر که شامل کلاس های ثابت و مشخصی بودند، لیبل های این دیتاست را تغییر و به شکل درستی به این مقادیر نسبت دادیم.

نامتوازن بودن داده‌ها: تفاوت بسیار زیاد در تعداد داده های کلاس، فرایند یادگیری را برای مدل سخت می‌کرد. با استفاده از تکنیک‌هایی نظیر data augmentation با استفاده از smote و مدل linearSVC که یادگیری خوبی در اسنگونه مسائل دارد، توانستیم عملکرد مدل را بهبود ببخشیم.

نیاز به پاکسازی داده‌ها: داده‌های دیتاست حاوی اطلاعات اضافه و بدون نیاز زیادی بوده(مانند تگ‌های html، علائم نگارشی، لینک‌هاو ...) که با تعریف تابع clean_text و استفاده از کتابخانه hazm ، این موارد را حذف کردیم.

انتخاب مدل و پارامتر مناسب: مدل LinearSVC به دلیل کارایی در داده‌های متنی انتخاب شد. تنظیمات شامل class_weight='balanced'برای رفع عدم تعادل، max_iter=10000و C=1.0 بود. با GridSearchCV و جستجوی[0.8, 1, 1.2]=C ، بهترین مقدار C=1.2 با دقت اعتبارسنجی91 به دست آمد. داده‌ها با نسبت 80-20 و stratify تقسیم شدند.

پیشنهادات:

با وجود SMOTE می‌توان از تکنیک‌های پیشرفته‌تر مانند تولید داده مصنوعی با مدل‌های زبانی مثل GPT برای فارسی، یا استفاده از نمونه‌برداری وزن‌دار در آموزش استفاده کرد تا تعادل بهتری ایجاد شود. همچنین جایگزینی LinearSVCبا مدل‌های مبتنی بر ترنسفورمر مانند ParsBERT می‌تواند با درک عمیق‌تر متن، دقت را افزایش دهد، هرچند نیازمند منابع محاسباتی بیشتری است.