به نام خدا

پدرام رستمی – ۸۱۰۱۰۰۳۵۳

گزارش تمرین دوم درس پردازش زبان طبیعی

در این پروژه سعی می‌کنیم دو طبقه‌بند برای مسئله‌های تشخیص اسپم بودن پیام‌ها و تشخیص دروغ بودن پیام بنویسیم. ساختار کلی هر دو طبقه‌بند تقریبا مانند همدیگر است. در هر دو مسئله، ابتدا داده‌ها پیش پردازش می‌شوند و سپس توکنایز شده و سپس بردار ویژگی‌ها را تشکیل داده و به وسیله‌ی بردارهای ورودی‌ها و مدل‌های یادگیری ماشین، مسئله را حل می‌کنیم.

در مسئله‌ی تشخیص اسپم بودن پیام‌ها، ابتدا خطوط دیگر پیام‌ها را با خط اولشان در دیتاست ترکیب می‌کنیم تا تمام متن پیام‌ها در یک ستون (ستون text) قرار بگیرد. سپس تمام \n ها را حذف می‌کنیم. در بخش پیش پردازش این مسئله، ابتدا تمام علائم نشانه گذاری به جز علائم $%?! را حذف می‌کنیم و قبل و بعد از علائم نشانه گذاری حذف نشده هم space اضافه می‌کنیم تا حتما به عنوان یک توکن جدا حساب شوند. برای نرمالایز کردن متن، تمام حروف را تبدیل به حروف کوچک می‌کنیم. همچنین تمام stop word ها را هم با استفاده از stop word های تعریف شده در کتابخانه‌ی nltk برای زبان انگلیسی حذف می‌کنیم. برای توکنایز کردن، از ابزار word_tokenize در کتابخانه‌ی nltk استفاده می‌کنیم. همچنین برای اینکه توکن‌هایمان بر اساس بن واژه‌ی کلمات باشد، از ابزار stemmer کتابخانه‌ی nltk استفاده می‌کنیم.

تمام مراحل پیش پردازش، از مرحله‌ی حذف علائم نشانه گذاری تا stem کردن توکن‌ها، را در تابع custom_analyzer قرار می‌دهیم. سپس این تابع را به عنوان analyzer به تابع TfidfVectorizer از کتابخانه‌ی sklearn می‌دهیم. از این تابع برای بدست آوردن بردارهای tfidf توکن‌ها استفاده می‌کنیم. این تابع شامل دو بخش مجزا برای پیش پردازش و توکنایز کردن کلمات است که هر دوی آن‌ها قابل overwrite کردن هستند و در صورتی که تابع analyzer به آن داده شود، از این تابع به جای هر دو بخش استفاده می‌کند. ویژگی‌هایی که TfidfVectorizer تولید می‌کند شامل بردارهای tfidf هر پیام است.

پس از به دست آوردن بردارهای ویژگی‌ها، از مدل svm در کتابخانه‌ی sklearn برای طبقه‌بندی بر اساس ویژگی‌ها استفاده می‌کنیم. نتایج این مدل مطابق جدول زیر است.

	Precision	Recall	F1-score	Support
Ham	1.00	0.98	0.99	992
spam	0.87	0.99	0.93	123
accuracy			0.98	1115
Macro avg	0.94	0.99	0.96	1115
Weighted avg	0.98	0.98	0.98	1115

همانطور که مشخص است، عملکرد مدل svm برای تشخیص پیام‌های spam از پیام‌های عادی بسیار مناسب است.

برای حل مسئله‌ی تشخیص دروغ بودن یا نبودن پیام هم تقریبا از ساختار مشابهی استفاده شده است. در این مسئله در بخش پیش پردازش، تمام علائم نشانه گذاری حذف شده است. همچنین برای استخراج ویژگی‌ها، به جز استفاده از بردارهای tfidf متن پیام‌ها، از فیلدهای anger، fear، joy، disgust و sad هم استفاده شده است. در این مسئله هم از طبقه بند svm استفاده شده است. نتایج این طبقه بند در جدول زیر قابل مشاهده است.

	Precision	Recall	F1-score	Support
False	0.59	0.39	0.47	553
True	0.63	0.79	0.70	714
accuracy			0.62	1267
Macro avg	0.61	0.59	0.58	1267
Weighted avg	0.61	0.62	0.60	1267

در نگاه اول، به نظر می‌آید که عملکرد طبقه بند مناسب نیست ولی نتایجی که برای مدل مقاله‌ی این مجموعه داده و به کمک BERT و دیگر ویژگی‌های این مجموعه داده طراحی شده است، دارای دقتی معادل 0.67 است. در نتیجه دقت مدل svm که بسیار ساده‌تر است و از ویژگی‌های کمتری هم استفاده می‌کند، تقریبا قابل قبول است.