محمدرضا احمدی تشنیزی
۹۸۱۷۰۶۴۶
پروژه یادگیری ماشین

۱.فهرست کتابخانه‌های مورد استفاده و دلایل آن:

الف) pandas :برای کار با مجموعه داده‌ها، از pandas استفاده می‌کنیم. این کتابخانه ساختارهای داده‌ای موثری مانند DataFrame را ارائه می‌دهد که امکان دستکاری و تحلیل داده‌ها را فراهم می‌کند.

ب) numpy :برای محاسبات عددی و عملیات آرایه، از numpy استفاده می‌شود.

ج) matplotlib و seaborn :این کتابخانه‌ها برای تصویرسازی داده‌ها استفاده می‌شوند. آن‌ها در ایجاد انواع نمودارها کمک می‌کنند، که به فهم توزیع و رفتار داده کمک می‌کند.

د) nltk و gensim : از این کتابخانه‌ها برای وظایف پردازش زبان طبیعی استفاده می‌شود. برای وظایف پیش‌پردازش متن مانند توکن‌سازی، حذف کلمات توقف، ساقه‌سازی و غیره، از NLTK استفاده می‌شود Gensim . برای ساخت مدل Word2Vec استفاده می‌شود.

ه) Sklearn : یک کتابخانه جامع برای وظایف یادگیری ماشین است. این کتابخانه شامل الگوریتم‌های مختلفی برای دسته‌بندی، رگرسیون، خوشه‌بندی و غیره است. همچنین ابزارهایی را برای پیش‌پردازش داده‌ها، اعتبارسنجی متقاطع، و تنظیم پارامترهای بیشینه‌سازی ارائه می‌دهد.

و) re :این یک ماژول داخلی پایتون برای کار با عبارات منظم است. در مرحله پیش‌پردازش برای حذف هر گونه کاراکتر ناخواسته از متن استفاده می‌شود.

۲.استخراج ویژگی در رویکرد دوم:(Word2Vec)

رویکرد دوم از مدل Word2Vec برای استخراج ویژگی استفاده می‌کند Word2Vec .یک مدل تعبیه کلمه آموزش دیده است که برای هر کلمه در متن، یک نمایش برداری ایجاد می‌کند. این بردارها، زمینه معنایی کلمات را در خود جای می‌دهند، جایی که کلمات مشابه بردارهایی دارند که در فضای برداری نزدیک‌تر هستند.

در این رویکرد، مراحل زیر برای استخراج ویژگی دنبال می‌شود:

الف) پیش‌پردازش: داده‌های متنی پیش‌پردازش می‌شوند تا تمام کلمات را به حروف کوچک تبدیل کنند، کاراکترهای خاص را حذف کنند، کلمات توقف را حذف کنند، و جملات را به کلمات توکن‌سازی کنند.

ب) آموزش مدل :Word2Vec پس از پیش‌پردازش، یک مدل Word2Vec بر روی متن آموزش دیده می‌شود. این مدل یک واژه‌نامه از متن آموزشی ایجاد می‌کند و سپس تعبیه کلمات را یاد می‌گیرد.

ج) ایجاد بردارهای عبارت: برای هر عبارت (یا جمله) در مجموعه داده‌ها، با میانگین گیری بردارهای Word2Vec تمام کلمات در عبارت، نمایش برداری را محاسبه می‌کنیم. اگر یک کلمه در واژه‌نامه Word2Vec نباشد، آن را نادیده می‌گیریم.

بردارهای عبارت حاصل به عنوان ویژگی‌ها برای وظیفه دسته‌بندی استفاده می‌شوند. باید توجه داشت که این ویژگی‌ها روابط معنایی بین کلمات را در خود جای می‌دهند، در مقابل رویکرد پایه ای کیسه‌ی کلمات.

۳.فهرست مدل‌های استفاده شده، هدف، و تعداد پارامترها

مدل: رگرسیون لجستیک

هدف: رگرسیون لجستیک به عنوان مدل اصلی برای وظیفه دسته‌بندی احساسات استفاده شد. این یک الگوریتم ساده اما قدرتمند برای مشکلات دسته‌بندی دودویی و چنددسته‌ای است. این مدل با استفاده از یک تابع لجستیک، احتمالات نتایج مختلف ممکن برای متغیر وابسته رده‌بندی را پیش‌بینی می‌کند.

تعداد پارامترها: در رگرسیون لجستیک، تعداد پارامترها به تعداد ویژگی‌ها بستگی دارد. در این مورد، وقتی از تعبیه کلمات به عنوان ویژگی استفاده می‌کنیم، تعداد پارامترها برابر خواهد بود با بعد تعبیه کلمات به علاوه یک (برای عبارت انتساب). برای نمونه، اگر از تعبیه 300 بعدی استفاده کنیم، آنگاه مدل رگرسیون لجستیک 301 پارامتر خواهد داشت.

اگر از یک مدل رگرسیون لجستیک با نظم‌دهی (که معمول است) استفاده کنیم، خواهد بود یک پارامتر اضافی Cکه کنترل معکوس قدرت نظم‌دهی را در دست دارد. این یک پارامتر در همان معنی با وزن و عبارت انتساب نیست (در طول آموزش به روز نمی‌شود)، اما باز هم بخش مهمی از پیکربندی مدل است.

۴.نتایج ارزیابی مدل‌ها

عملکرد مدل‌ها می‌تواند بر اساس معیارهای دسته‌بندی مانند دقت، دقت، بازخوانی و امتیاز F1 ارزیابی شود.

رویکرد 1: روش‌های اولیه

دقت مدل رگرسیون لجستیک با ویژگی‌های TF-IDF تقریبا 0.55 بود. دقت میانگین وزن‌دار، بازخوانی، و امتیاز F1به ترتیب 0.51، 0.55، و 0.47 بودند.

رویکرد Word2Vec2 :

دقت مدل رگرسیون لجستیک با ویژگی‌های Word2Vec تقریبا 0.51 بود. دقت میانگین وزن‌دار، بازخوانی، و امتیاز F1 به ترتیب 0.44، 0.51، و 0.41 بودند.

در هر دو رویکرد، مدل بر روی احساس خنثی (برچسب 2) که بیشترین نمونه‌ها در مجموعه داده را داشت، عملکرد بهتری داشت. این نشان‌دهنده نیاز به رسیدگی به نامتوازنی کلاس در مجموعه داده است، زیرا مدل به سمت پیش‌بینی کلاس اکثریت سوق داده شده است. تکنیک‌هایی مانند بیش‌نمونه‌سازی کلاس اقلیت، کم‌نمونه‌سازی کلاس اکثریت، یا استفاده از وزن‌های کلاس می‌توانند برای رفع این مشکل استفاده شوند.

5. تحلیل نقاط قوت و ضعف مدل

نقاط قوت:

سادگی و کارایی: رگرسیون لجستیک مدلی نسبتاً ساده است که سریع آموزش داده می‌شود و پیش‌بینی انجام می‌دهد، که این موضوع آن را گزینه خوبی برای مدل‌های پایه می‌کند.

تفسیر احتمالاتی: مدل‌های رگرسیون لجستیک احتمالات مرتبط با هر کلاس را ارائه می‌دهند به جای پیش‌بینی‌های سخت. این در شرایطی مفید است که می‌خواهیم درجه اطمینان مدل را بدانیم.

اهمیت ویژگی: مدل‌های رگرسیون لجستیک می‌توانند بینش‌هایی درباره اهمیت ویژگی ارائه دهند، که می‌تواند به عنوان سهم هر ویژگی در پیش‌بینی تفسیر شود.

نقاط ضعف:

خطی بودن: رگرسیون لجستیک حد فاصل تصمیم خطی را فرض می‌کند، که می‌تواند با داده‌های پیچیده که حد فاصل تصمیم خطی نیست، محدود کننده باشد.

نقطه‌های پرت: رگرسیون لجستیک می‌تواند به نقطه‌های پرت حساس باشد. یک نقطه خارج از محدوده در فضای ویژگی می‌تواند حد فاصل تصمیم را به طور چشمگیری تغییر دهد.

چند خطی بودن: رگرسیون لجستیک می‌تواند با چند خطی بودن مشکل داشته باشد، جایی که یک ویژگی پیش‌بینی می‌تواند به صورت خطی از دیگران پیش‌بینی شود.

6. مقایسه رویکرد اول و دوم

در رویکرد اول، ما از TF-IDF به عنوان روش استخراج ویژگی استفاده کردیم. این روش تنها فرکانس یک کلمه را در یک سند خاص (یا جمله، در این زمینه) در نظر می‌گیرد، بلکه نیز نسبت معکوس آن کلمه در کل مجموعه را در نظر می‌گیرد.

در رویکرد دوم، ما از Word2Vec برای استخراج ویژگی استفاده کردیم. Word2Vec، بر خلاف TF-IDF، می‌تواند شباهت معنایی کلمات را به دست آورد چرا که این مدل نمایش کلمات را از متن در جملات یاد می‌گیرد.

با این حال، در آزمایشات ما، رویکرد اول (TF-IDF) کمی عملکرد بهتری نسبت به رویکرد دوم (Word2Vec) داشت. این می‌تواند به این دلیل باشد که رگرسیون لجستیک ممکن است بهترین مدل برای گرفتن روابط معنایی و دستوری که Word2Vec یاد می‌گیرد، نباشد. همچنین ممکن است مدل Word2Vec بر روی مجموعه کافی یا مرتبط بزرگ آموزش دیده نشده باشد تا روابط معنایی معنی‌دار بین کلمات را به دست آورد.