دانشکدگان علوم
دانشکده ریاضی، آمار و علوم کامپیوتر

پروژه حسابگری زیستی مهلت تحویل: ۲۰ تیر

تحلیل احساسات متن فارسی

در این پروژه می‌خواهیم با تحلیل احساسات بر روی متن فارسی استخراج شده از توییتر آشنا شویم.تحلیل احساسات شامل شناسایی و دسته‌بندی احساسات یا نظرات افراد مرتبط با موضوعات مختلف است. . برای آشنایی بیشتر این مقاله را مطالعه کنید.

مجموعه دادگان

مجموعه دادگان را دریافت کنید. توجه داشته باشید که این مجموعه داده با داده مورد استفاده در مقاله متفاوت است.فقط ستون‌های tweet و emotion برای این سوال مورد نیاز هستند. کلاس‌های موجوددر ستون emotion و تعداد نمونه‌های هر کلاس را به کمک یک نمودار میله‌ای نمایش دهید.

پیش‌پردازش داده‌ها

پیش‌پردازش متن در پردازش زبان طبیعی برای بهبود عملکرد مدل بسیار مهم است.مراحل پیش‌پردازش ذکر شده در قسمت ۳.۲ مقاله را اعمال کنید و برای هر یک از مراحل، یک مثال که پیش‌پردازش مورد نظر روی آن اعمال شده است را قبل و بعد از پیش‌پردازش چاپ کنید. توجه داشته باشید ممکن است برخی از این مراحل برای این مجموعه داده نیاز نباشد و تغییری در هیچ یک از سطرها ایجاد نکند. برای انجام پیش‌پردازش‌های این بخش می‌توانید از کتابخانه re و کتابخانه‌هایی که پیش‌پردازش زبان فارسی را پشتیبانی می‌کنند استفاده کنید

نمایش ویژگی

در وظایف پردازش زبان طبیعی ، داده هایی‌که به طور کلی پردازش می‌شوند، متن خام هستند.با این حال، مدل‌ها فقط میتوانندداعداد (Id) را پردازش کنند، بنابراین بایداز توکن‌سازها برای تبدیل متن خام به اعداداستفاده کنید.داده‌های متن پیش‌پردازش شده را با توکن‌ساز ParsBERT به اعداد تبدیل کنید. برای این که تمام سطرها طول یکسانی داشته باشند، از

padding استفاده کنید و حداکثر طول جملات را برابر با ۳۲ در نظر بگیرید. Embedding کلمات را به عنوان بردارها در فضایی با ابعاد بالا نشان می‌دهند که روابط معنایی را به تصویر می‌کشند. این تعبیه‌ها مدل‌های یادگیری ماشین را قادر می‌سازندالگوها و احساسات را از داده‌های متنی بیاموزند.در این مرحله، به کمک مدل از پیش آموزش دیده ParsBERT بردار تعبیه را برای ورودی‌ها به دست آورید. با تغییر Configuration مدل، ابعاد بردار تعبیه را برابر با ۱۲۰ در نظر بگیرید. توجه داشته باشید که برای مدیریت حافظه می‌توانید از تکنیک‌هایی مانند تکه‌تکه کردن و کتابخانه gc استفاده کنید. ابعاد پیش‌فرض بردارتعبیه در ParsBERT چقدر است؟تعداد ابعاد این بردار بیانگر چیست؟مفهوم بردارتعبیه را توضیح دهید و بیان کنید به نظر شما کدام یک از کلمات موجود در مجموعه داده ممکن است تعبیه نزدیک به هم داشته باشند؟

ساخت مدل

داده‌ها را با نسبت۷۰-۳۰به دو دسته آموزش و تست تقسیم کنید و۰.۲ از داده‌های آموزش را به عنوان اعتبارسنجی در نظر بگیرید. الگوریتم جستجوی حریصانه برای یافتن هایپرپارامترهای بهینه برای مدل CNN-LSTM را در فضای جستجو با ۸حالت زیر اعمال کنید و در نهایت هایپرپارامترهای بهینه که منجر به کمترین خطای اعتبارسنجی می‌شود را گزارش کنید. هایپرپارامترهای دیگر مدل را مطابق با جدول ۳ مقاله اعمال کنید.

```
batch_sizes = [8, 64]
learning_rates = [0.001, 0.0001]
optimizers = [Adam, SGD]
```

۳ در مرحله بعد، مدل‌های CNNو LSTM ساده را با هایپرپارامترهای بهینه به دست آمده ایجاد کرده و آموزش دهید. نیازی به اعمال الگوریتم جستجوی حریصانه برای این مدلها نیست. به نظر شما، هر یک از این مدل‌ها چه نقاط ضعف و چه نقاط قوتی دارند و ادغام این دو مدل با چه هدفی انجام میشود؟

ارزیابی

داده‌های تست را به کمک معیارهای ارزیابی ذکر شده در مقاله ارزیابی کنید و یک جدول مشابه جدول ۴مقاله برای مدلهایCNNو LSTM LSTM-CNN چاپ کنید. روش‌های macro averaging، micro averaging و weighted averaging برای محاسبه میانگین معیارهای ارزیابی را مقایسه کنید و توضیح دهید هر یک از این روشها چه تاثیری بر مقدار عددی این معیارها در این مسئله دارد.

امتیازی

از روش Bag of Words برای نمایش ویژگی استفاده کنید و به کمک کتابخانه sklearn روش‌های سنتی ماشین لرنینگ که در مقاله ذکر شده‌اند را آموزش داده و به کمک دادگان تست ارزیابی کنید. نتایج را به جدول نتایج بخش قبل اضافه کنید. برای کاهش استفاده از منابع، در این بخش میتوانید از بخشی از داده‌ها نمونه گرفته و از این نمونه‌ها برای آموزش و ارزیابی مدل‌ها استفاده کنید.