گزارش پروژه – کیمیا صدیقی – 97243046

Part 1:

در ابتدا کتابخانه ی hazm را نصب می کنیم.

Part 2:

هر آنچه که نیاز است import شود را import می کنیم، از جمله کتابخانه هایی مثل Tensorflow، nltk ، sklearn، numpy، pandas و ... .

Part 3:

سپس در این مرحله دیتاست خود را می خوانیم و آن را در متغیری مثل df ذخیره می کنیم.

Part 4,5:

پس از آن به مرحله preprocessing داده ها می رسیم که یکی از مراحل بسیار مهم در زمینه nlp می باشد. برای این کار ابتدا ردیف هایی از دیتاست را که خالی هستند و حاوی اطلاعات نمی باشند پاک می کنیم. پس از آن، هر entry از داده ها را به کمک word_tokenize به مجموعه ای از کلمات می شکنیم. سپس برای تمیز کردن داده ها، عباراتی مثل \n، \t و \r را از داده هایمان پاک می کنیم و به جای آن ' ' یا space قرار می دهیم. در ادامه از ماژول هایی مثل normalizer، stemmer و lemmatizer از کتابخانه ی hazm که نصب کردیم، برای تمیز کردن بیشتر داده ها استفاده می کنیم که normalizer مواردی مثل وجود فاصله و نیم فاصله را کنترل می کند. stemmer کلمه ای مانند کتاب ها را به کتاب تبدیل می کند. و lemmatizer افعالی مثل می روم را به رفت تبدیل می کند.

Part 6:

در این قسمت داده ها را به دو قسمت training set و test set تقسیم می کنیم.

Part 7:

در این مرحله هایپر پارامترهایی مثل vocab_size و embedding_dim که در مرحله ی ساخت مدل استفاده می شود و trunc_type، padding_type و oov_tok که در مرحله ی vectorize کردن استفاده می شود را تعیین می کنیم.

Part 8:

این مرحله vectorize کردن داده ها می باشد. که خود شامل دو مرحله fit_on_texts و texts_to_sequences می باشد. که مرحله fit_on_texts دیکشنری کلمات را بر اساس متن داده ها به روز می کند. در این روش ایندکس هر واژه به تکرار آن در

متن ها وابسته است. به این معنا که هر چه شماره خانه ی یک کامه کوچکتر باشد، تعداد تکرار آن در متن ها بیشتر است. در مرحله text_to_sequences هر کلمه را در هر متن می گیرد و شماره خانه ی مربوط به آن واژه در دیکشنری را جایگزین آن می کند.

Part 9:

چون دیتاست ما به صورت categorical می باشد پس label هر متن و خود هر متن که در مرحله قبل vectorize شده است را، به روش one_hot انکود می کنیم.

Part 10:

مرحله بعد ساخت مدل می باشد که از ساختار CNN استفاده کرده ایم که دارای لایه هایی از جمله Embedding، Conv1D، GlobalMaxPooling1D، Dropout، Flatten و Dense می باشد. که در Conv1D از هایپر پارامترهایی که در چند مرحله پیش تعیین شد استفاده می شود. vocab_size تعداد بیشترین کلماتی که باید استفاده شوند را بیان می کند که این کلمات پر کاربردترین کلمه ها در داده های ما هستند. پس از آن تعداد گام ها(epoch) و batch_size را مشخص کرده و مدل را بر روی داده هایمان fit می کنیم.

Part 11:

در این قسمت روند کارایی و بهبود یا عدم بهبود مدل در هر گام را بر روی نمودار میبینیم.

Part 12:

در این قسمت برای خروجی گرفتن فایل test مراحل preprocess را که در مرحله ترین شدن مدل، بر روی دیتاست انجام دادیم، روی داده های فایل test نیز انجام میدهیم و سپس با استفاده از مدل آموزش داده شده، خروجی آن را predict می کنیم. و در نهایت label خروجی هر متن را دیکود کرده تا به حالت اصلی خود برگردد و نهایتا خروجی را در یک فایل با پسوند csv ذخیره می کنیم.