پردازش زبان‌های طبیعی

تمرین سوم

محمد جواد رنجبر

۸۱۰۱۰۱۱۷۳

بهار ۱۴۰۲

Contents

توضیح فایل‌ها:

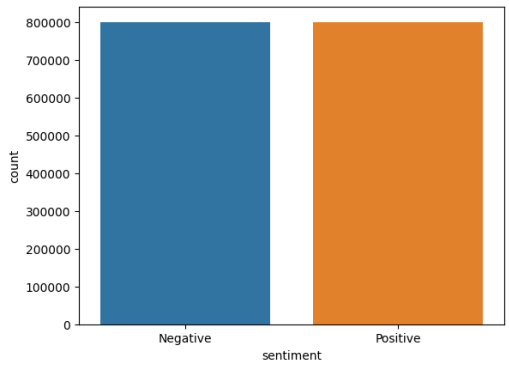
به واسطه سنگین شدن فایل colab تمرین در دو فایل انجام شده است که در فایل CA3_Q_one_hot.ipynb بخش‌های وابسته به one hot encoding و در فایل CA2_glove_w2v.ipynb بقیه بخش‌های تمرین قرار داده شده است.

بخش ۱:

ابتدا پیش‌پردازش‌های مناسب برای این توییت‌ها را انجام می‌دهیم که شامل موارد زیر می‌باشد:

- حذف stop wordها
- حذف لینک‌ها و علامت‌های اضافی مانند # و @ (user ها و هشتگ نگه داشته می‌شوند).
- از آنجا که emojiها برای وظیفه، sentiment می‌توانند کاربردی باشند. آن‌ها را با اسامی emoji جایگزین می‌کنیم.
- Lemmetazation
- تبدیل lableها از ۰ و ۴ به ۰ و ۱ برای آسانتر شدن پردازش
- برای حل مشکل هم اندازه بودن توییت‌ها دو روش وجود دارد:
  - استفاده از padding برای توییت‌ها که در کد پیاده‌سازی شده است اما سرعت آموزش را به شدت پایین می‌آورد.
  - کنار هم گذاشتن توییت‌های هم اندازه در هر batch

همچنین نمودار توزیع داده‌های مختلف به شکل زیر می‌باشد، که تعداد داده‌ها در هر دوکلاس تقریبا برابر می‌باشد.

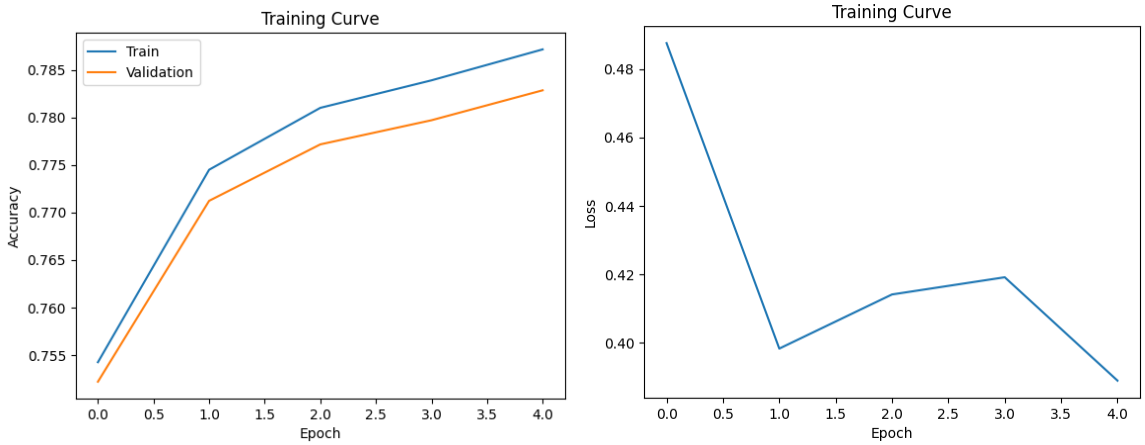


شکل ۱ توزیع داده‌ها

برای اینکه بتوانیم از داده‌ها استفاده کنیم و دچار مشکلات حافظه نشویم از ۱٫۶ میلیون توییت تعداد ۶۰۰۰۰ را به صورت تصادفی انتخاب می‌کنیم. حال با استفاده از white space کلمات را tokenize می‌کنیم و با استفاده از embeddingهای مختلف، مدل RNN را آموزش می‌دهیم.

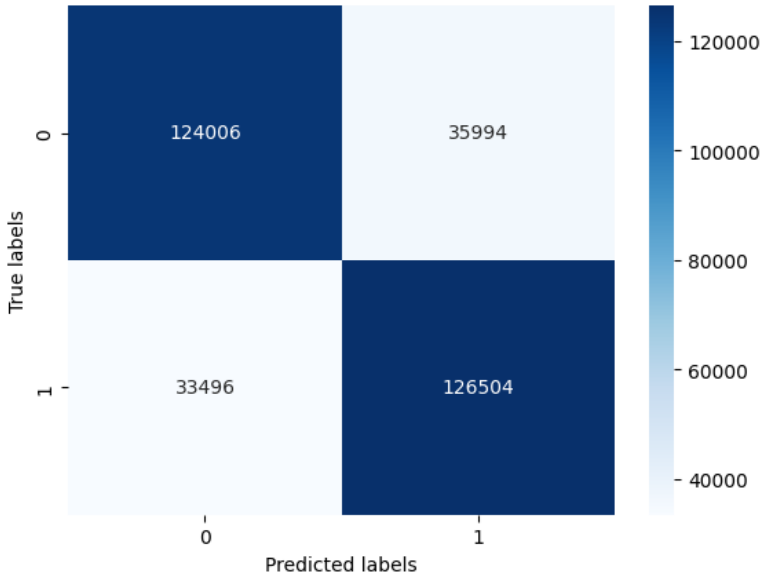
One Hot encoding:

این نوع embedding بسیار سنگین بوده و اصلا مناسب نیست، با این حال نتایج مدل به شکل زیر خواهد بود:



شکل ۲ عملکرد rnn برای one hot

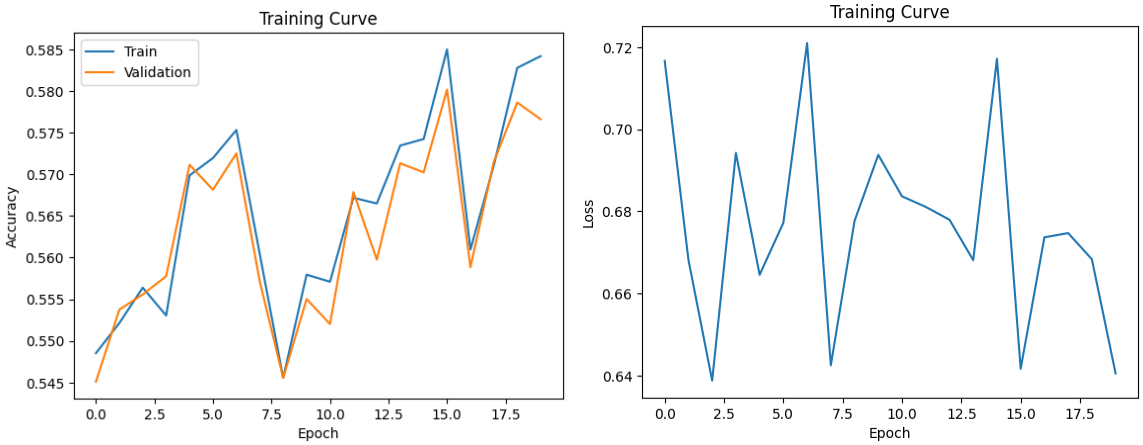
همچنین ماتریس درهمریختگی به شکل زیر می‌باشد که بیشتر داده‌ها به درستی دسته‌بندی شده‌اند و مدل اشتباه خاصی نکرده است.



شکل ۳ ماتریس درهمریختگی برای مدل rnn و one hot

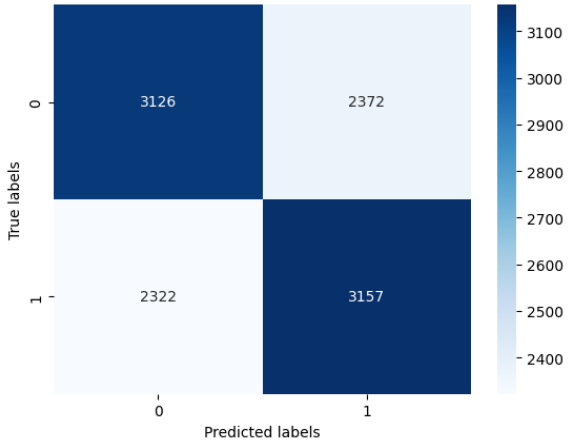
Glove:

نتایج آموزش برای مدل RNN با این embedding به صورت زیر است.



شکل ٤ عملکرد مدل rnn برای glove

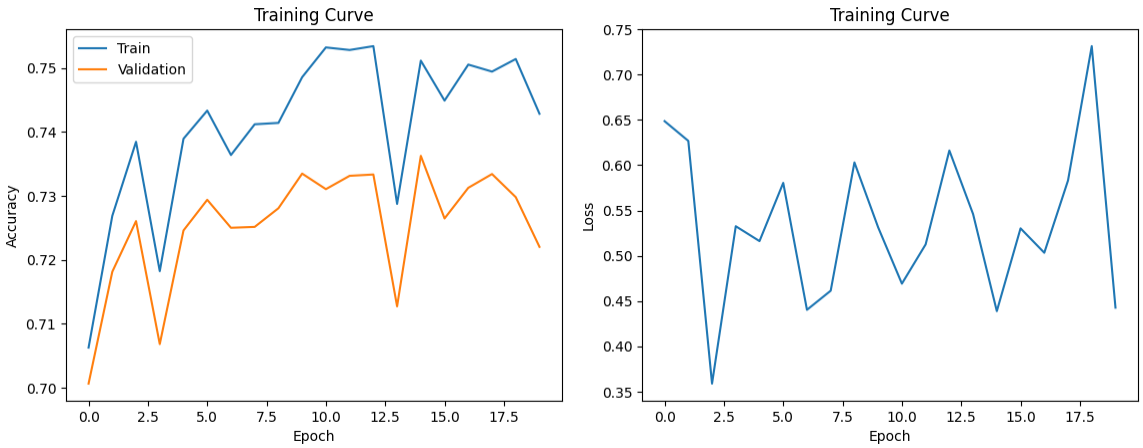
همانطور که مشخص است مدل دقت آنچنان خوبی ندارد و می‌تواند تعداد دفعات بیشتری نیز آموزش یابد. همچنین ماتریس درهمریختگی این مدل به شکل زیر می‌باشد:



شکل ٥ ماتریس درهمریختگی مدل RNN با glove

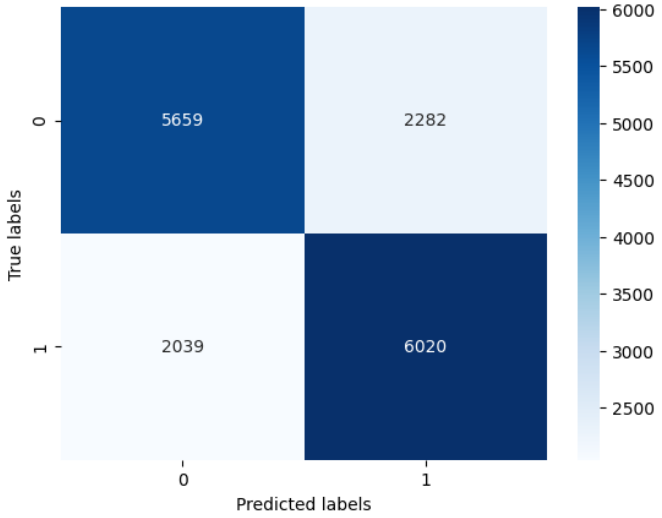
Word۲vec:

این embedding نسبت به دو embedding دیگر بهتر عمل می‌کند و نتایج آن به صورت زیر است.



شکل ۶ عملکرد مدل rnn برای word2vec

همچنین ماتریس درهمریختگی به شکل زیر می‌باشد:



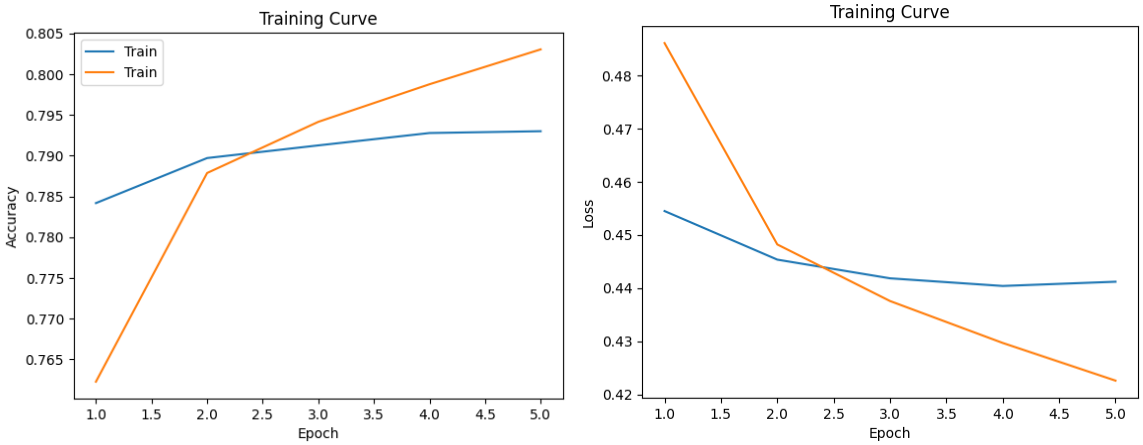
شکل ۷ ماتریس درهمریختگی مدل rnn برای word2vec

بخش ۲:

LSTM:

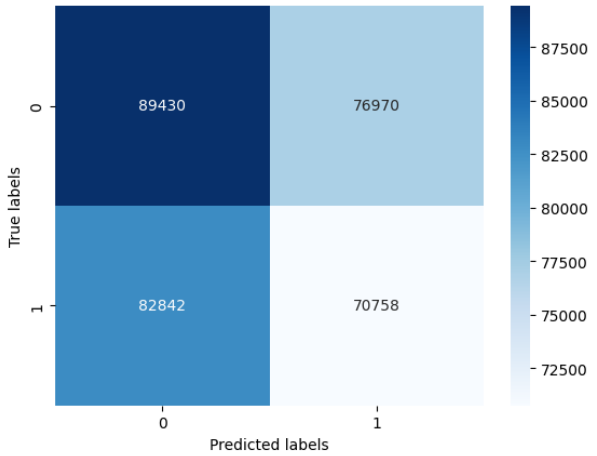
One Hot encoding:

نتایج آموزش برای این مدل به شکل زیر می‌باشد:



شکل ۸عملکرد lstm برای one hot

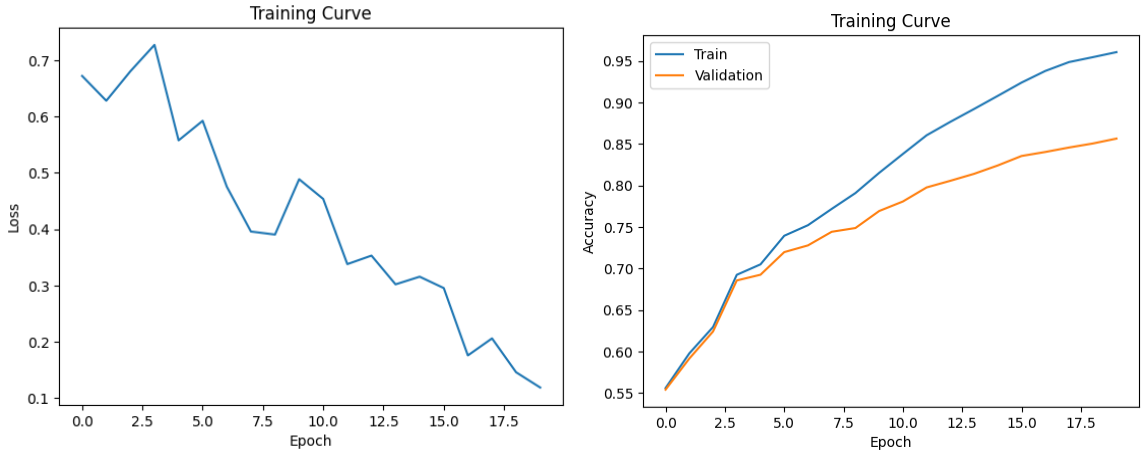
همچنین ماتریس در همریختگی برای این مدل به شکل زیر خواهد شد:



شکل ۹ ماتریس درهمریختگی برای مدل lstm و one hot

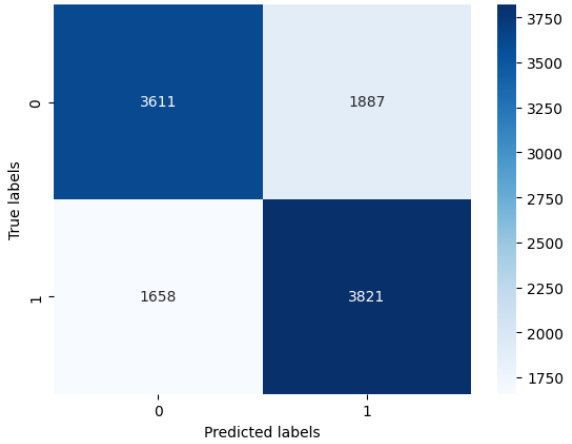
Glove:

نتایج مدل



شکل ۱۰ عملکرد lstm برای glove

همچنین ماتریس درهمریختگی این مدل به شکل زیر می‌باشد:



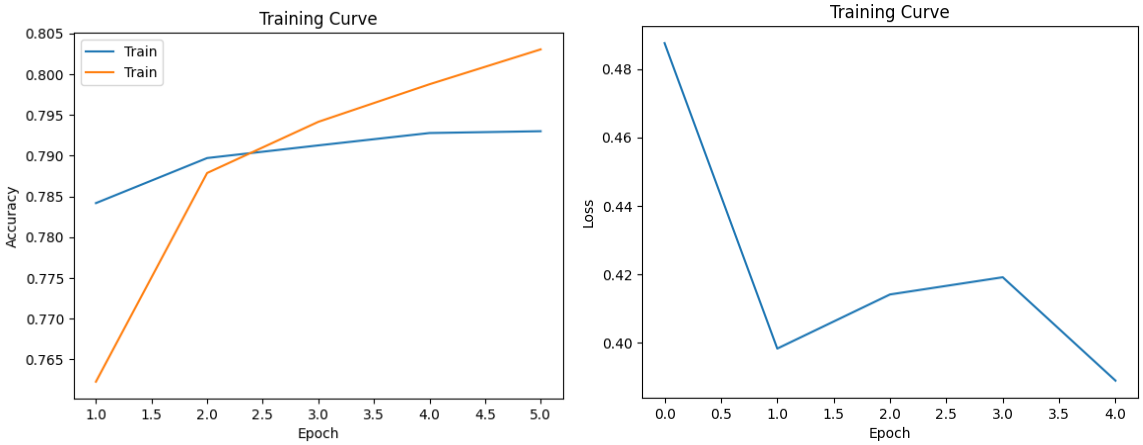
شکل ۱۱ ماتریس درهمریختگی LSMT با glove

این مدل به شدت بهتر از مدل RNN عمل می‌کند که با توجه به پیچیده‌تر بودن آن همین نیز انتظار می‌رفت.

GRU:

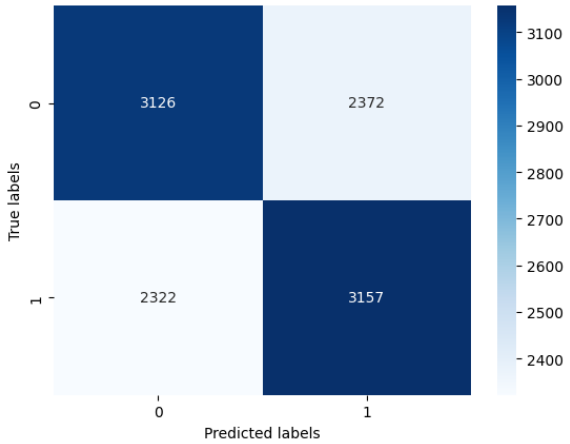
One Hot encoding:

عملکرد مدل به شکل زیر است:



شکل ۱۲ عملکرد مدل gru برای one hot

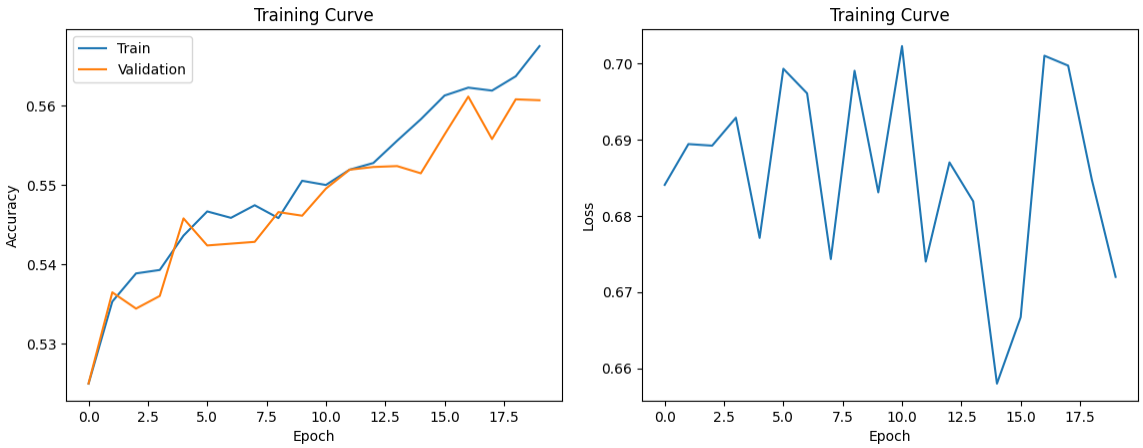
ماتریس درهمریختگی به شکل زیر می‌باشد:



شکل ۱۳ ماتریس درهمریختگی مدل gru و one hot

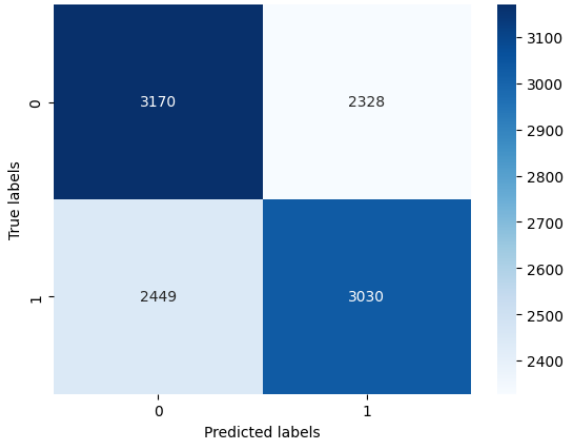
Glove:

نتیجه حاصل از این مدل به شکل زیر می‌باشد:



شکل ۱۴ عملکرد gru برای glove

همچنین ماتریس در همریختگی این مدل به شکل زیر می‌باشد:



شکل ۱۵ ماتریس درهمریختگی gru با glove

مقایسه دو مدل:

همانطور که پیش‌بینی می‌شد مدل GRU به نسبت سریعتر از مدل LSTM می‌باشد. اما این سریعتر بودن باعث کمی ضعیف‌تر عمل کردن این مدل نسبت به مدل LSTM نیز می‌شود که مورد انتظار بود. به طور کلی مدل lstm بهتر عمل کرد اما کمی کندتر نیز بود.