بسمه تعالی



**موضوع**: فاز نهایی پروژه پردازش زبان طبیعی

**نام استاد**: دکتر مینایی

**دانشجو**: محمد صدرا خاموشی‌فر- رضا علیدوست

دانشکده مهندسی کامپیوتر

بهار **1401**

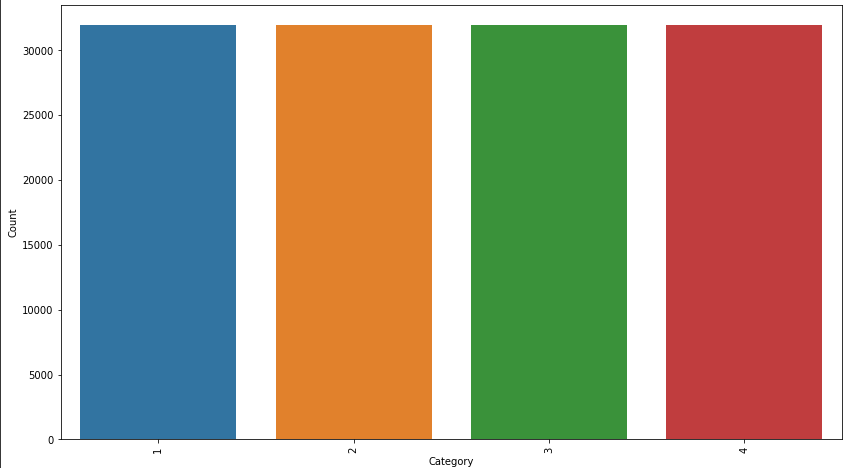
**توجه داشته باشید که تمامی توضیحات مربوط به کد در داخل فایل نوت بوک‌ها موجود است در این گزارش فقط نتایج و مقایسه آورده شده است.**

# بخش دوم

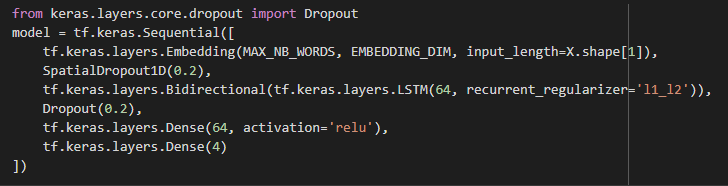
در این بخش سعی کردیم از یک مدل baseline برای شروع کار استفاده کنیم. همچنین از سه نوع متفاوت دیتا ست مختلف استفاده کردیم تا مدل طراحی شده را بسنجیم.

## داده های Ag news

این داده ها داده های سبکی بوده اند که در سایت Kaggle وجود دار که لینک آن در نوت بوک موجود است. این داده ها به صورت یکنواخت و هم اندازه توزیع شده اند و در 4 دسته‌ی مختلف قرار گرفته اند.



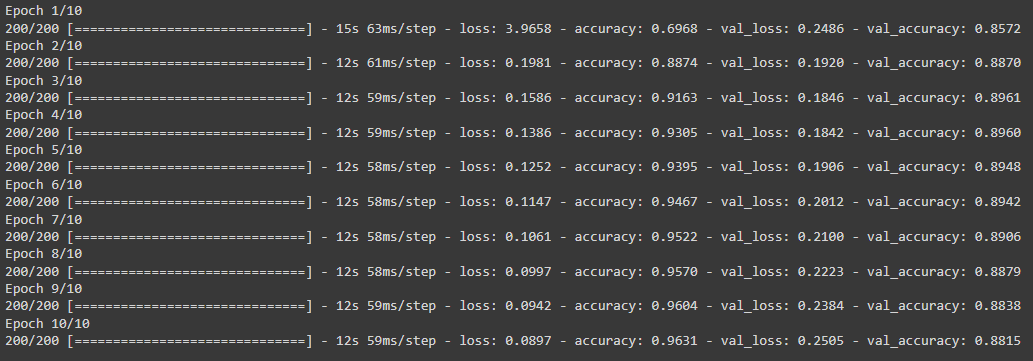
بعد از آموزش دادن آن روی مدل زیر :



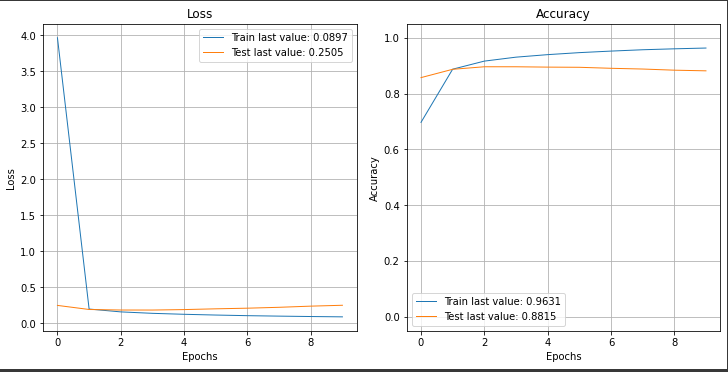
و همچنین برای هاپر پارامتر ها هم از هایپر پارامتر های زیر استفاده کردیم:



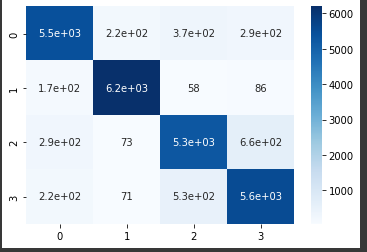
نتایج آن به صورت زیر شد »



مدل خیلی overfit نشده است و همچنین رو داده های ولیدیشن عملکرد خوبی داشته و به دقت تقریبا 90 رسیده.

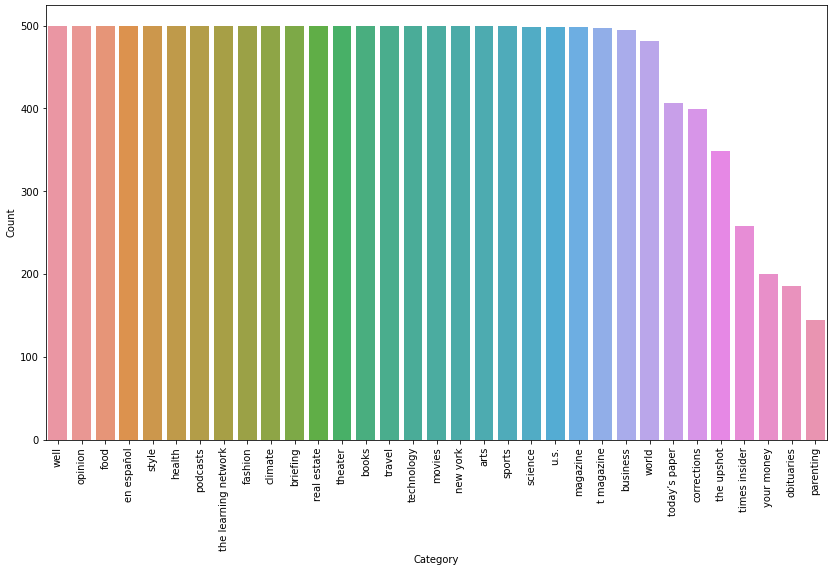


همه این ها به دلیل این است که شاید تسک جدا کردن 4 تا کلاس با استفاده از 120000 کار خیلی دشواری نبوده است.



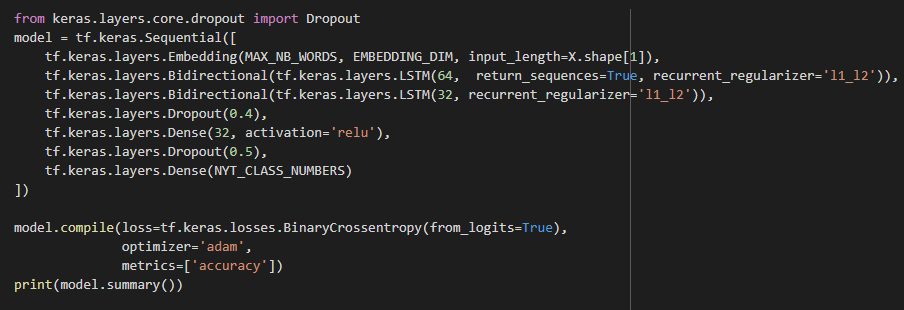
## داده های New York Times news

این داده ها همان داده های خودمان است که در فاز اول پروژه تولید کرده بویدیم و شامل اخبار حاصل از 33 کتگوری مختلف میباشد. اما به دلیل محدودیت هایی که داشتیم متاسفانه نشد که داده های زیادی جمع آوری کنیم و در نهایت از هر category نهایتا 500 تا داده داشتیم. توزیع داده ها به صورت شکل زیر میباشد:

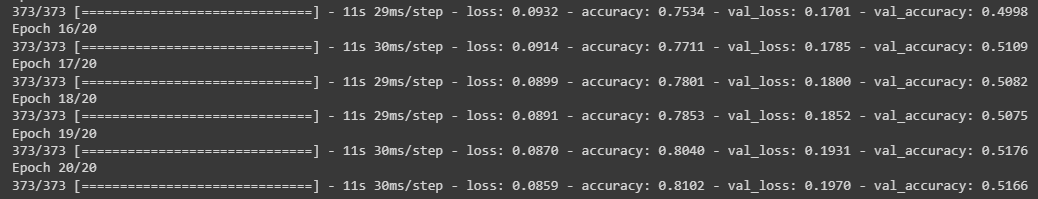


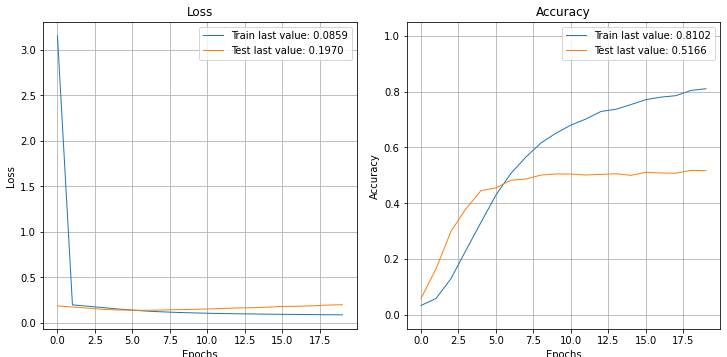
همانطور که مشاهده میکنید از برخی از داده ها حتی کمتر از 200 داده داریم. این ها باعث میشود که بتوان پیشبینی کرد که مدل به احتمال زیاد overfit خواهد شد و یا اصلا مدل چیزی یاد نمیگیرد.

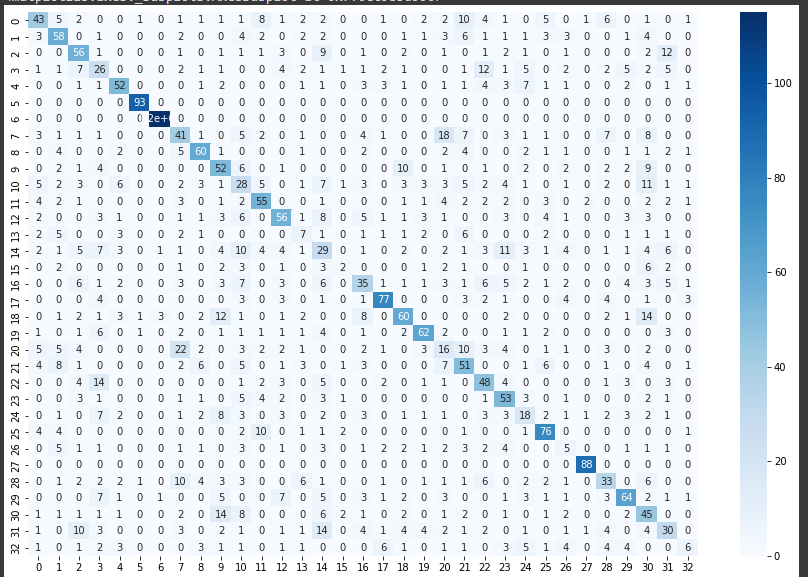
با استفاده از روش ها regulizer l1 + l2 و همچنین استفاده از لایه های Dropout خواستیم تا حدودی جلوی overfit شدن مدل زیر را بگیریم :



اما پس از آموزش مشاهده شد که همانطور که پیش بینی میشد، مدل overfit شده است و دقت روی داده های validation از یه حدی بیشتر بالا نمیرود.



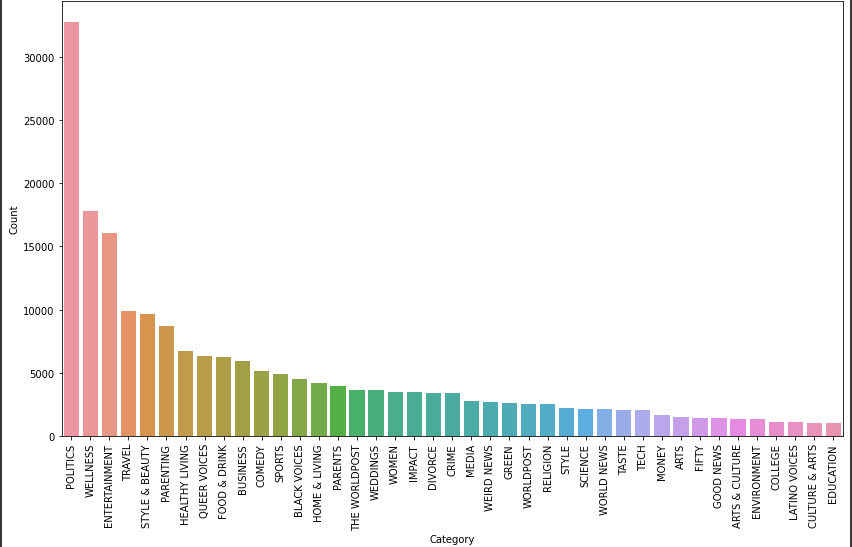




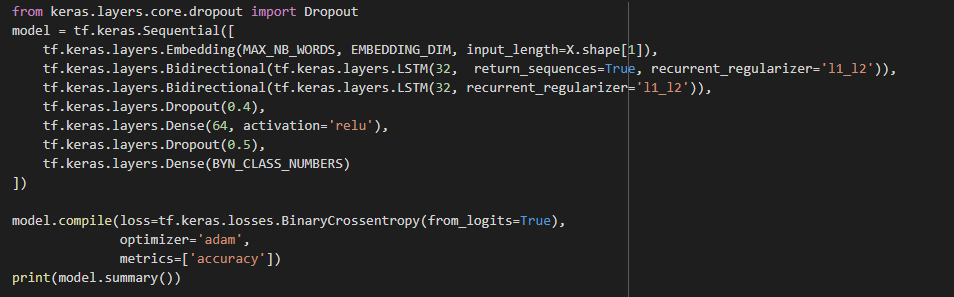
## داده های Breaking news

در نهایت از این داده ها استفاده کرده تا بتوانیم روی داده های اخبار واقعی و به اندازه کافی به یک دقت قابل قبولی برسیم. داده های breaking news نیز از سایت Kaggle برداشته شده است که بسیار معروف بوده و در موارد زیادی مورد استفاده قرار گرفته شده.

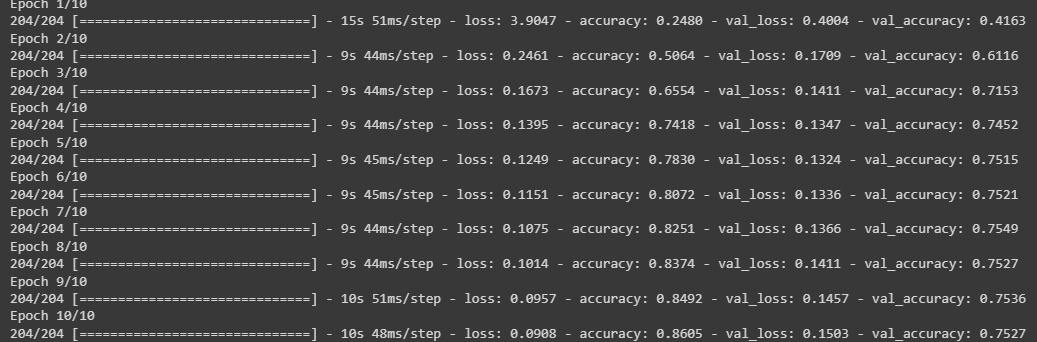
توزیع این داده ها را در شکل زیر میتوانید مشاهده کنید :



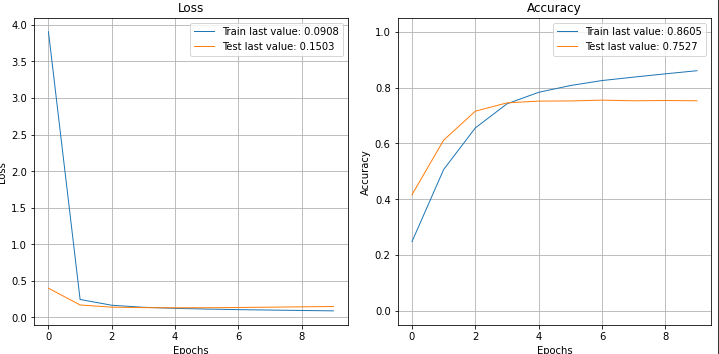
همانطور که مشاهده میشود این داده ها هم توزیع یکسانی ندارند اما تعداد آن ها برای آموزش یک مدل deep تقریبا کافی است. برای همین ما 12 تا کتگوری ای که داده ی بیشتر داشته است را جدا کردیم و مدل خودمان را رو آن ها ترین کردیم. تعداد کل آن داده ها تقریبا برابر 130000 مقدار میباشد :

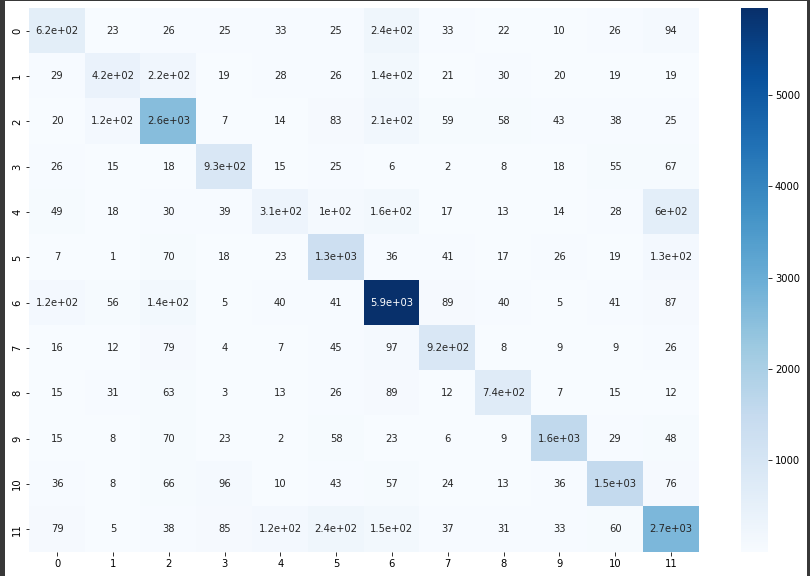


از مدل بالا برای آموزش روی داده های breaking news استفاده کردیم و زیاد تغییری نسبت به حالت قبلی نداشته است . اما نتایح آن بهتر از حالت قبل میشود. بعد از آموزش دقت آن روی داده هایی که تا حالا ندیده است به تقریبا 75 درصد میرسد اما همچنان مدل تاحدی overfit شده است :



میتوان نتایج آن را روی نمودار هم مشاهده کرد:





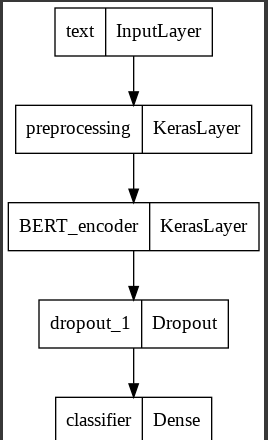
بدلیل یکنواخت نبودن داده ها این امر اتفاق افتاده است که میتوان با کارهایی نظیر کاهش دادن داده های زیاد و یا data augmentation تا حدی اثر این مشکل را برطرف نمود. در حالت کلی برای تسک های text classification دیتا خیلی خیلی مهم است به طوری که یک دیتای خیلی مرتب و اندازه تقریبا رو تمامی مدل ها performance یکسانی دارد به همین دلیل به دلیل کمبود دیتای های new York times نتوانستیم دقت خوبی کسب کنیم.

# بخش اول و سوم(مدل BERT)

برای تمامی موارد بالا مدل برت را آموزش داده ایم و به نتایجی شبیه همان نتایج بالا رسیدیم:

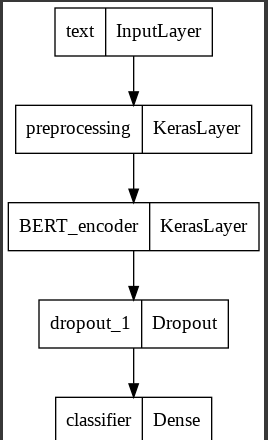
## داده های Ag news

مدل استفاده شده :



نتایج داخل تصویر بسیار ریز بود. در داخل نوت بوک میتوانید مشاهده کنید.

## داده های New York Times news



نتایج مدل BERT بسیار شبیه به حالت های ساده قبل میشد به دلیل اینکه تسک text classification به کار های دیگری همچون text preprocessing و تعداد داده و توزیع مناسب آنها بستگی دارد و آن چنان به پیچیدگی مدل بستگی ندارد و این باعث میشود که نتایح به هم شبیه باشند. اما در تمامی حالت ها مدل bert متناظر عملکرد بهتری داشته اسنت.