گزارش نهایی پروژه

علی علی‌محمّدی ۹۶۱۳۰۲۷

استفاده از شبکه‌های عصبی برای دسته‌بندی متون مربوط به COVID-19

مقدمه

تحلیل احساسات[1] همواره در طی ادوار مختلف از اهمیت بالایی در جامعه‌ی انسانی برخوردار بوده است؛ زیرا انسان موجودی اجتماعی‌ست و نظرات دیگر اعضای اجتماع بر زندگی او تأثیرگذار است. در بسیاری از مسائل دنیای مدرن کنونی نیز دانستن احساس دیگران نسبت به یک پدیده اهمیت خاص خود را دارد؛ مثلاً مهم است که مردم یک کشور رئیس‌جمهور خود را قبول دارند یا خیر؛ یا مثلاً خریداران این محصول از آن رضایت دارند یا خیر. تمام این اطلاعات می‌تواند برای بهبود وضعیت (بالا بردن محبوبیت رئیس‌جمهور یا بهبود کیفیت محصول) به کار گرفته شود.

در این پروژه سعی شده است که با استفاده از تکنیک‌های دسته‌بندی متن[2]، توییت‌های مربوط به بیماری COVID-19 را با به کار گیری رویکرد تحلیل احساسات، با بهره‌گیری از ابزار شبکه‌ی عصبی دسته‌بندی و برچسب‌گذاری کنیم.(Koyel Chakraborty, 2020)

اهمیت این موضوع در آن‌جا مشخص می‌شود که می‌توان با استفاده از نتایج به دست آمده از این دسته‌بندی و Meta-Dataی مربوط به توییت‌ها تشخیص داد که مثلاً در کدام یک از مناطق جغرافیایی، اخباری مثبت درباره‌ی ویروس پخش شده است و یا جو روانی حاکم بر یک ناحیه به چه صورت است. آیا همه‌ی توییت‌ها بسیار منفی هستند و در آن ناحیه تعداد زیاد ابتلا یا مرگ‌ومیر باعث این نتیجه شده است؟ آیا عده‌ای جواب روانی بسیار منفی را ایجاد می‌کنند یا گسترش می‌دهند که روحیه‌ی جامعه را تضعیف می‌کند؟ اخبار امیدبخش کدامند؟ با توجه به اطلاعات به دست آمده، در کدام جامعه نیاز است تا تبلیغات بیشتری برای ترغیب افراد آن جامعه به رعایت پروتکل‌های بهداشتی انجام شود؟

[1] Sentiment Analysis
[2] Text Classification

مجموعه‌ی داده‌ها

از مجموعه‌ی داده‌های Coronavirus tweets NLP - Text Classification موجود در Kaggle برای این منظور استفاده شده است که از قبل برچسب‌گذاری شده‌اند.

نحوه‌ی ارزیابی

با توجه به این که برای این منظور از شبکه‌های عصبی استفاده می‌شود، ملاک ارزیابی، دقت[3] و مقدار تابع Loss خواهد بود.

پیش‌نیازها

برای انجام این پروژه از کتاب‌خانه‌های استاندارد Python و همچنین کتاب‌خانه‌ی یادگیری عمیق Keras (TensorFlow) استفاده شده است. برای پیش‌پردازش متن از کتاب‌خانه‌ی استاندارد re و کتاب‌خانه‌ی پردازش زبان طبیعی NLTK استفاده شده است. برای انجام محاسبات از کتاب‌خانه‌ی NumPy و برای کار با فایل‌ها و داده‌های ورودی از کتاب‌خانه‌ی Pandas بهره گرفته شده است. برای دیداری‌سازی و رسم نمودارها و شکل‌های مختلف از کتاب‌خانه‌های PIL، matplotlib، plotly، Seaborn و WordCloud استفاده شده است. برای محاسبه‌ی معیارهای سنجش مدل‌ها و متریک‌ها[4]ی مربوط به آن‌ها، برداری‌سازی متن[5] و استفاده از دسته‌بندها[6]ی پرکاربرد از کتاب‌خانه‌ی sklearn استفاده شده است.

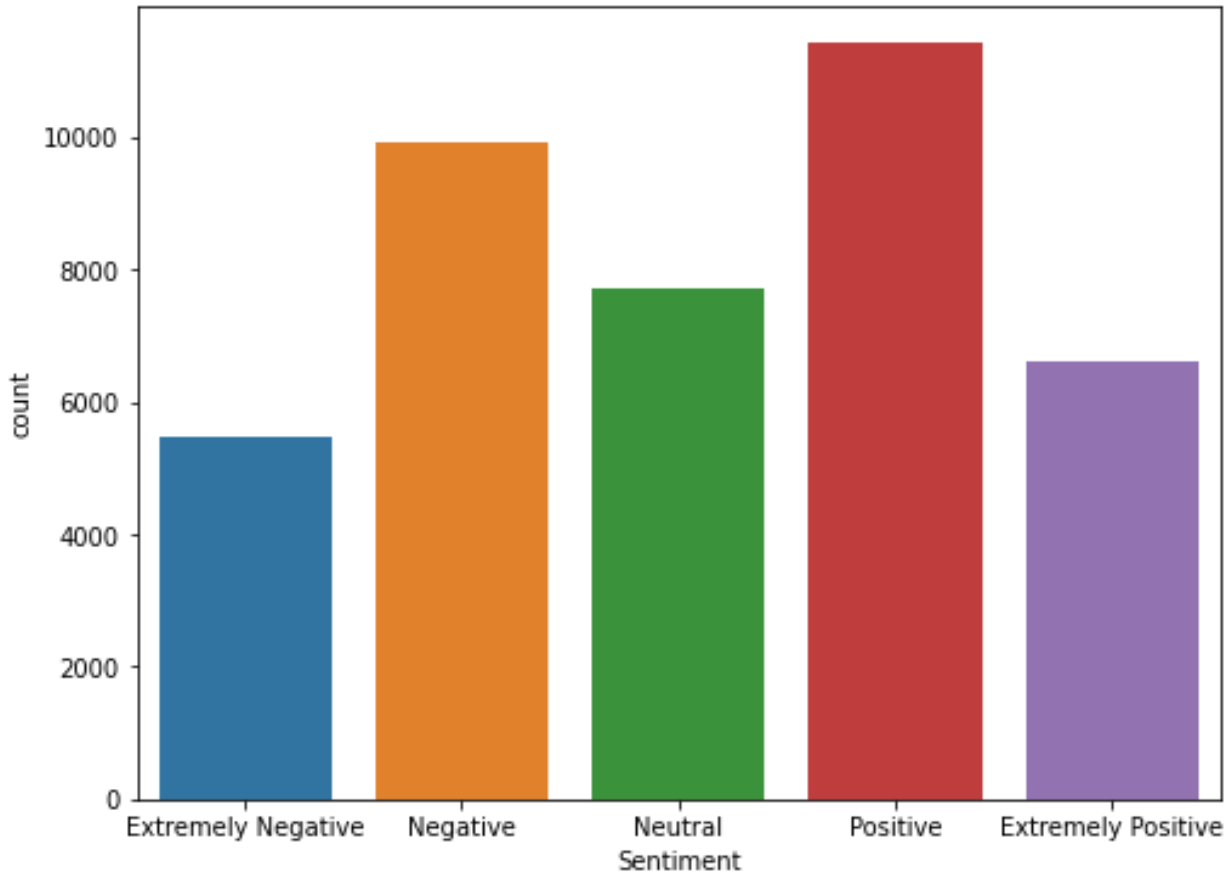
دقت شود که برای این که توانایی بازسازی شرایط مشابه را در اجراهای متفاوت برنامه داشته باشیم، شماره‌ی دانشجویی را به عنوان seed به تابع random داده‌ایم:

```
np.random.seed(9613027)
```

[3] Accuracy
[4] Metrics
[5] Vectorization
[6] Classifiers

شرح کار

در ابتدا یک نمودار «تعداد توییت بر حسب برچسب» با استفاده از Seaborn رسم می‌شود تا دید کلی نسبت به داده‌ها به دست آید.

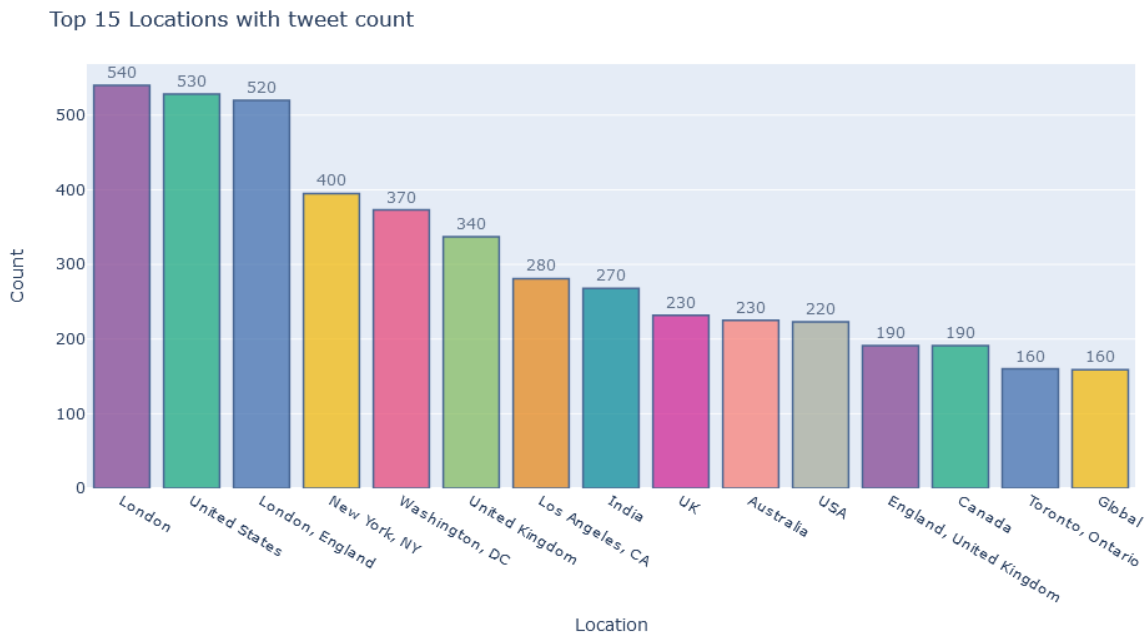


همان‌طور که مشاهده می‌شود، بیشترین توییت‌ها مربوط به دسته‌ی Positive و کمترین توییت‌ها مربوط به دسته‌ی Extremely Negative هستند.

مشخصات کلی مجموعه‌ی داده‌های آموزش به شرح زیر است:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41157 entries, 0 to 41156
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   UserName       41157 non-null  int64 
 1   ScreenName     41157 non-null  int64 
 2   Location       32567 non-null  object
 3   TweetAt        41157 non-null  object
 4   OriginalTweet  41157 non-null  object
 5   Sentiment      41157 non-null  object
dtypes: int64(2), object(4)
memory usage: 1.9+ MB
```

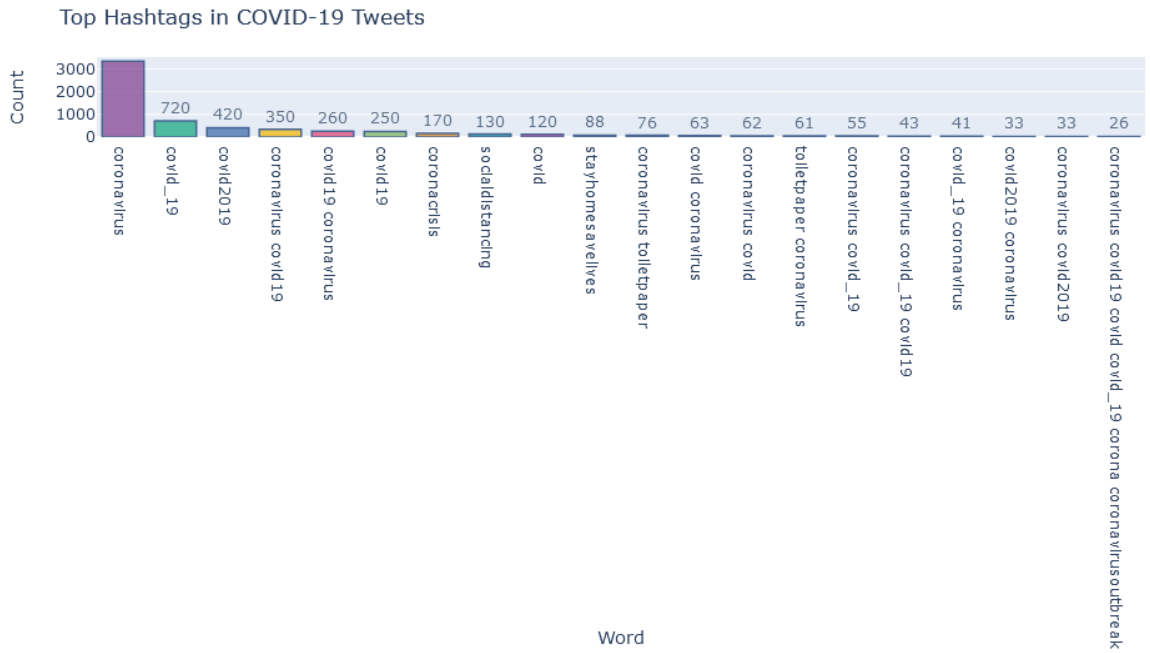
از داده‌های بالا نتیجه گرفته می‌شود که ۸۵۹۰ توییت داریم که Location مشخصی ندارند. با شمارش توییت‌های مربوط به هر موقعیت جغرافیایی نمودار زیر به دست می‌آید:



✓ توجه کنید که نمودار فوق به صورت interactive در notebook مربوط به پروژه موجود است.

با تشخیص هشتگ‌ها[7] می‌توان مشاهده کرد که بیشترین هشتگ‌های استفاده شده به طور مستقیم به کلمات "Corona" یا "COVID" مربوطند که طبیعی است:

Word	Count
coronavirus	3354
covid_19	723
covid2019	420
coronavirus covid19	349
covid19 coronavirus	257



✓ توجه کنید که نمودار فوق به صورت interactive در notebook مربوط به پروژه موجود است.

[7] Hashtags

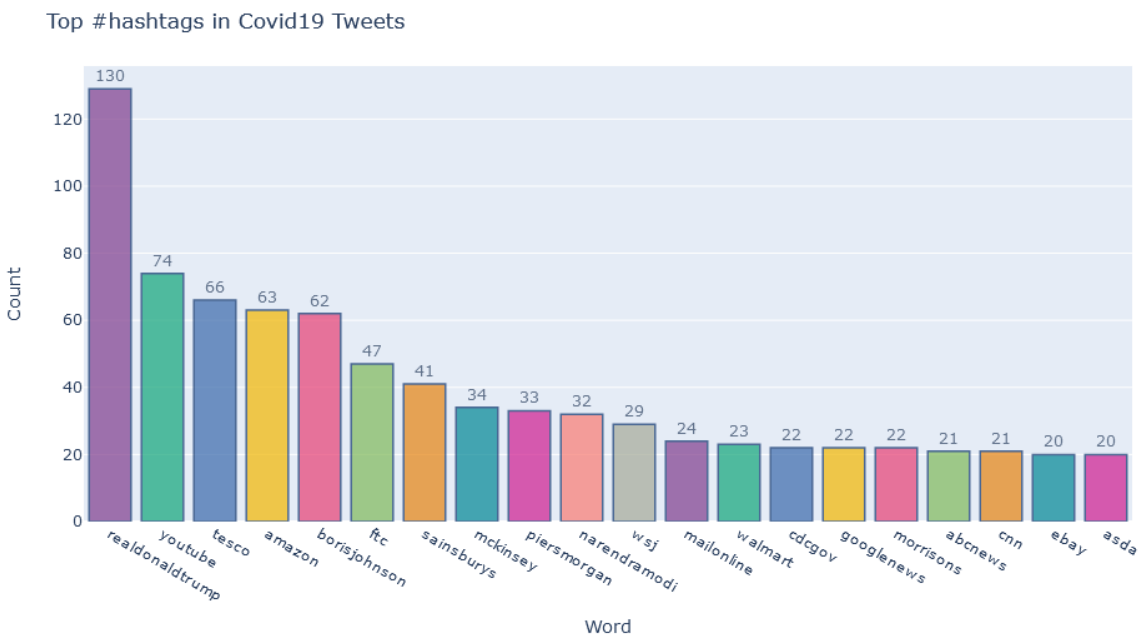
با استخراج mentionها از توییت‌ها در می‌یابیم که اکانت تأیید شده‌ی رئیس‌جمهور وقت ایالات متحده‌ی آمریکا، دونالد ترامپ، با حساب کاربری realdonaldtrump با اختلاف بسیار، در بحث‌ها وجود داشته است که با توجه به سخنان و سیاست‌های جنجالی او، قابل توجیه است و مردم انتقادات زیادی از او داشتند.

در این میان، نام شرکت آمازون نیز به چشم می‌خورد که دلایل مختلفی از جمله ثروتمندتر شدن جف بزوس، مؤسس و مدیرعامل آن شرکت و مرگ عده‌ای از کارمندان این شرکت به دلیل عدم رعایت موارد بهداشتی که به بیماری COVID-19 مبتلا شدند، می‌توانند این موضوع را توجیه کنند.

نام شرکت YouTube نیز به این دلیل در بین پرکاربردترین‌ها قرار دارد که با اعمال قرنطینه در سراسر جهان، اقبال عمومی به سرویس‌های این شرکت جهش بزرگی داشت و حتی به مرحله‌ای رسید که اتحادیه‌ی اروپا در طی توافقی با YouTube، از آن شرکت خواست تا کیفیت ویدئوهای استریم‌شده در اروپا را کاهش دهد تا از ترافیک بسیار سنگین شبکه و زیرساخت‌های اینترنت اروپا که به دلیل استفاده‌ی فزاینده‌ی مردم، افزایش چشم‌گیری داشت، کاهش یابد و شبکه از دسترس خارج نشود.

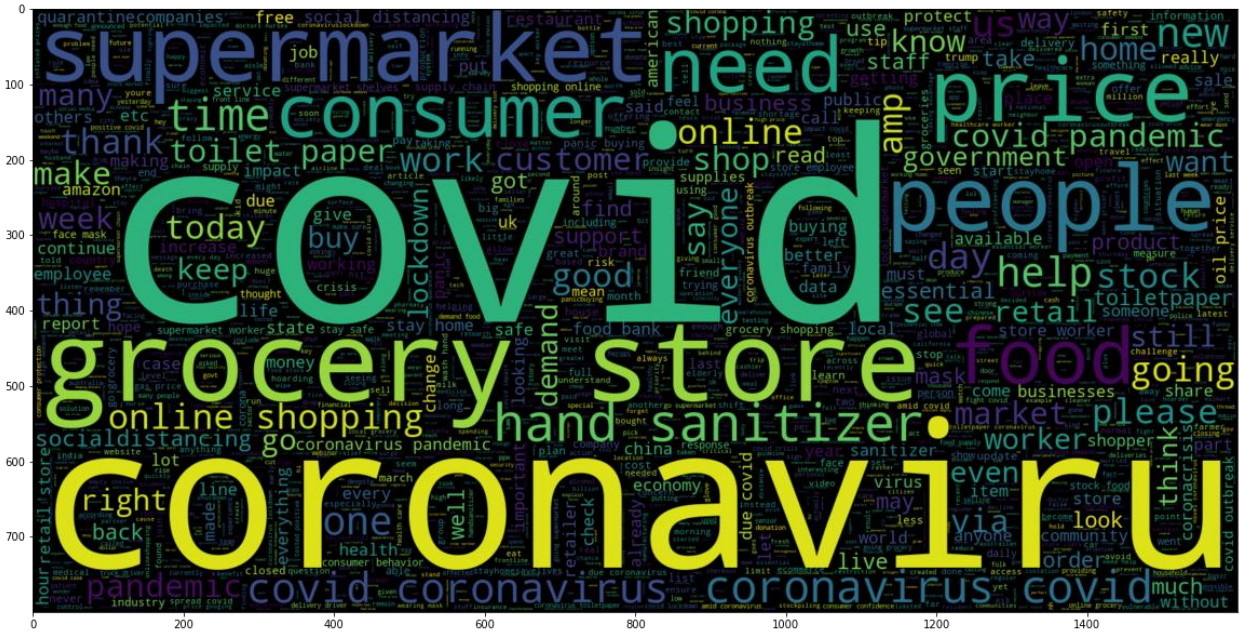
نام بوریس جانسون، نخست‌وزیر انگلیس نیز در این میان دیده می‌شود که به دلیل اصرار او به بی‌خطر بودن این ویروس و دست دادن با دیگران، و سپس بیمار شدن و تا پای مرگ پیش رفتن ایشان است که باعث شد تا ایشان موضع خود را نسبت به بیماری تغییر دهند و به خطرناک بودن آن پی ببرند.

Word	Count
realdonaldtrump	129
youtube	74
tesco	66
amazon	63
borisjohnson	62

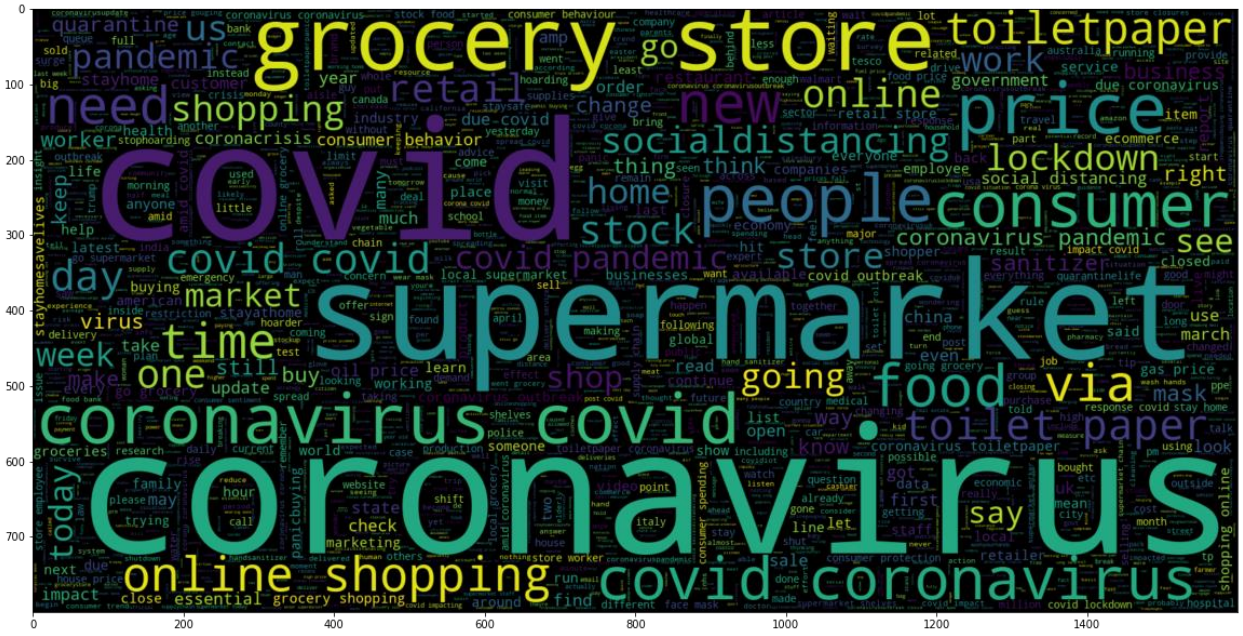


✓ توجه کنید که نمودار فوق به صورت interactive در notebook مربوط به پروژه موجود است.

اَبرِ کلمات[8] مربوط به احساسات مثبت:

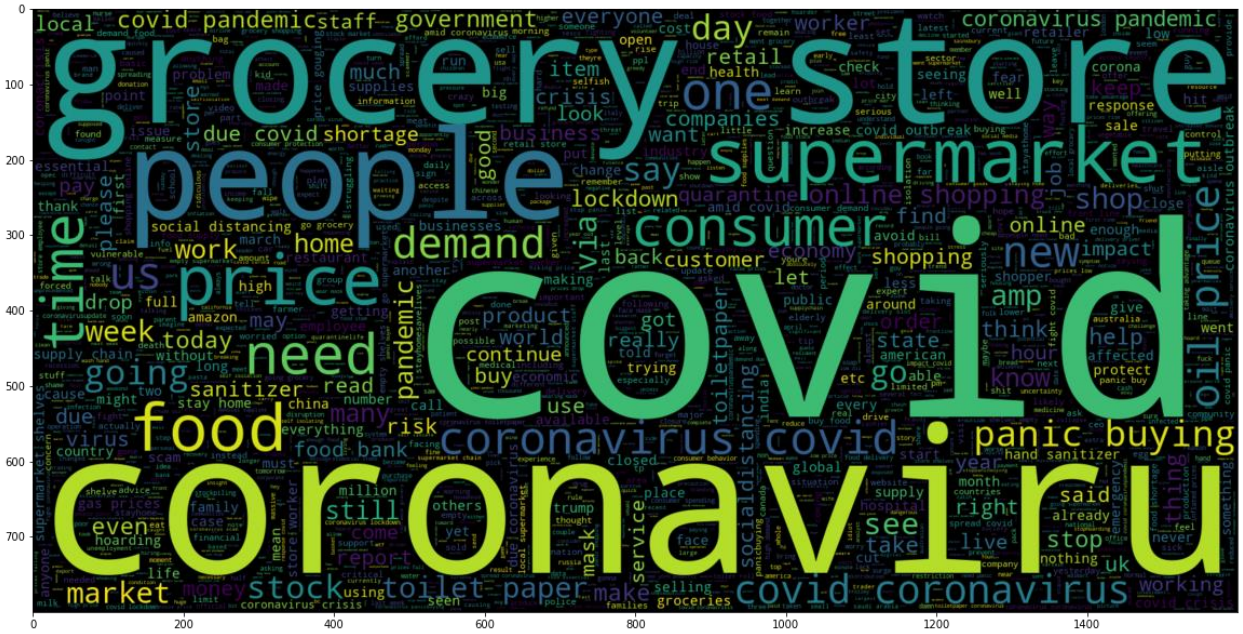


اَبرِ کلمات مربوط به احساسات خنثی:



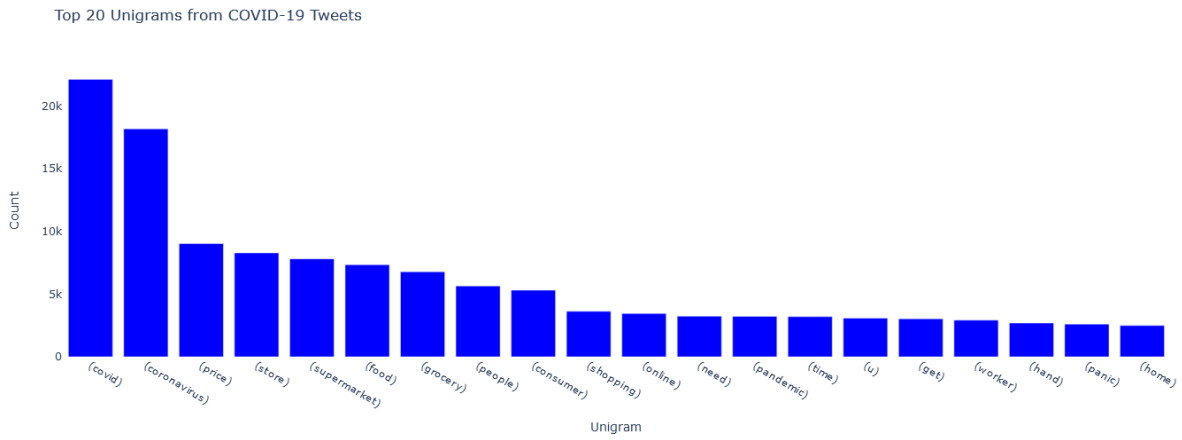
[8] Word Cloud

اَبرِ کلمات مربوط به احساسات منفی:



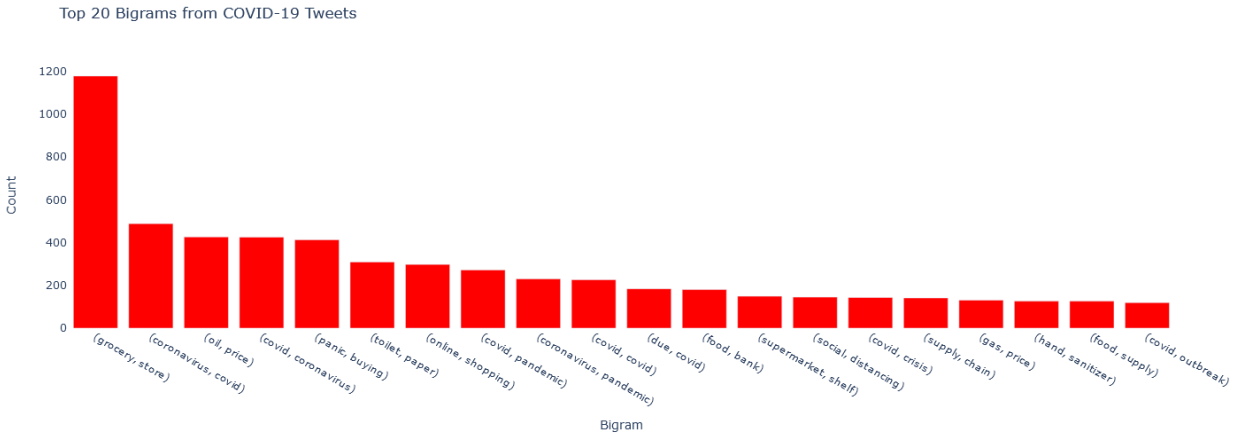
با تحلیل اصطلاحات پرکاربرد با استفاده از بررسی N-gramها می‌توان به نتایج جالبی دست یافت:

برای Unigramها همان‌طور که انتظار می‌رفت، کلمات "COVID" و "Coronavirus" بیشترین استفاده را داشته‌اند اما در وهله‌ی بعد، مسائل ناشی از کرونا دیده می‌شود که باعث نگرانی مردم شده است؛ مسائلی که مربوط به خرید مایحتاج روزانه‌ی زندگی است؛ مثل خرید از سوپرمارکت، مواد غذایی و قیمت‌ها.



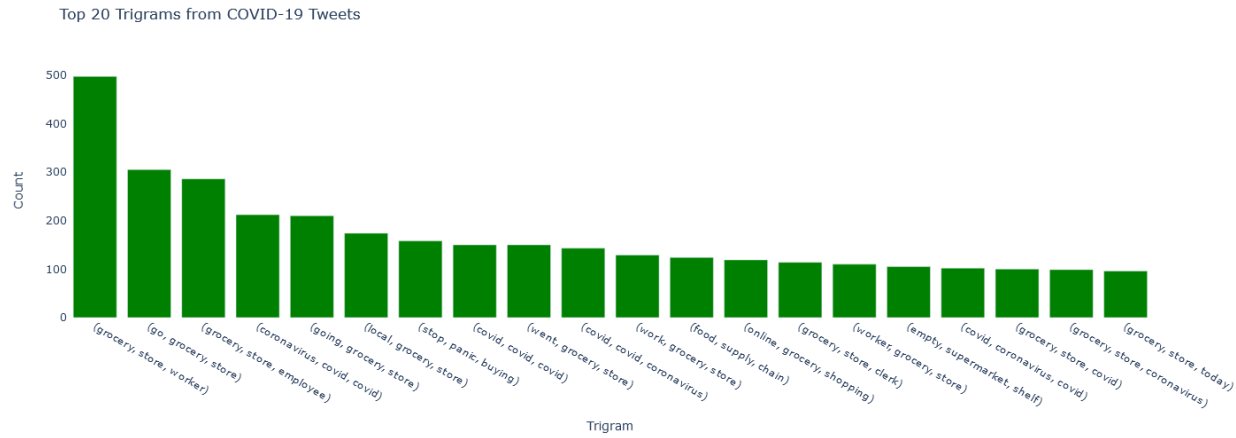
✓ توجه کنید که نمودار فوق به صورت interactive در notebook مربوط به پروژه موجود است.

با تحلیل Bigramها به طور دقیق‌تر به مسائل مورد توجه مردم پرداخته می‌شود. همان‌طور که مشاهده می‌شود، مردم درباره‌ی قیمت نفت، خرید مایحتاج از مغازه، خرید آنلاین، فاصله‌گذاری اجتماعی، خرید از روی وحشت و دستمال توالت به بحث می‌پرداخته‌اند.



✓ توجه کنید که نمودار فوق به صورت interactive در notebook مربوط به پروژه موجود است.

با بررسی Trigramها به اهمیت بسیار بالای خرید مایحتاج روزانه پی برده می‌شود؛ مشاهده می‌شود که هفت Trigram که در صدر لیست قرار دارند، همگی شامل کلمات grocery و store هستند. نکته‌ی جالب توجه این است که پرکاربردترین Trigram، grocery store worker است؛ این نشان‌دهنده‌ی اهمیت بسیار بالای شغل این دسته از افراد است. با توجه به این که سوپرمارکت‌ها جزو دسته‌ی مشاغل ضروری قرار دارند، کارمندان آن‌ها باید حتماً در محل کار خود حاضر شوند. این باعث شد تعدادی از این کارمندان به دلیل ابتلا به ویروس کرونا، جان خود را از دست بدهند که همین موضوع باعث شده است این موضوع در بین پربحث‌ترین‌ها قرار بگیرد.



✓ توجه کنید که نمودار فوق به صورت interactive در notebook مربوط به پروژه موجود است.

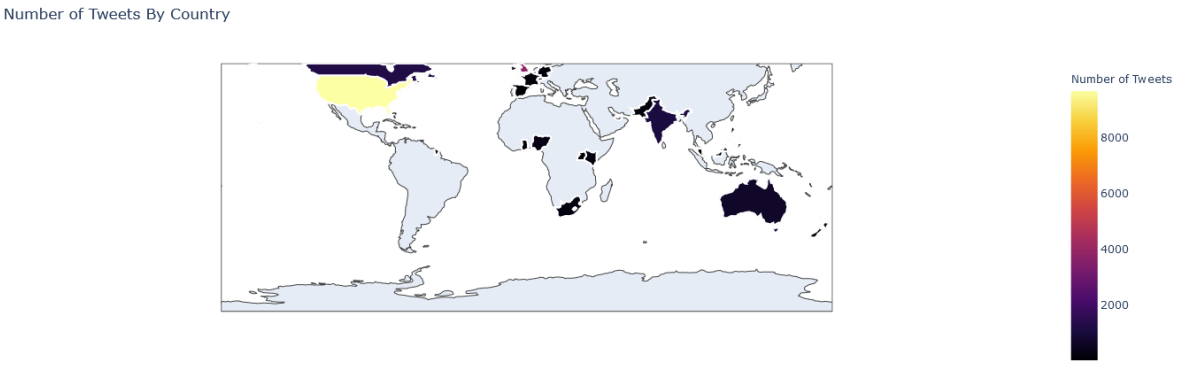
حال که به تحلیل کلی داده‌ها پرداخته شد، به سراغ ساخت مدل می‌رویم.

همان‌طور که در نمودارهای فوق مشاهده شد، در تحلیل موقعیت جغرافیایی می‌توانیم بهتر عمل کنیم. به همین دلیل، فیلترهایی تعریف کردیم که شهرهای یک کشور را در همان کشور در نظر بگیرد و نه به عنوان یک کشور. (همچنین اسم شهرها را تصحیح می‌کنیم.)

در نهایت، مشخصات داده‌های جغرافیایی به شرح زیر است:

```
None                    8579
London                   539
United States            526
London, England          520
New York, NY             395
                         ...
crystal palace             1
Callander Ontario          1
San Diego and beyond       1
 Emerald Isle ??           1
World Wide?                1
Name: Location, Length: 12212, dtype: int64
```

حال با استفاده از نمودار interactive، آن را روی نقشه‌ی جهان اعمال می‌کنیم:



✓ توجه کنید که نمودار فوق به صورت interactive در notebook مربوط به پروژه موجود است.

اکثر قریب به اتفاق این توییت‌ها از کشورهای انگلیسی زبان ارسال می‌شوند، که منطقی به نظر می‌رسد؛ زیرا این توییت‌ها همگی به زبان انگلیسی هستند. بزرگترین نقش در ارسال این توییت‌ها را ایالات متحده آمریکا و پس از آن، انگلیس و کانادا انجام داده‌اند.

تکرار اصطلاحات "grocery store"، "price"، "supermarket" و "online shopping" در توییت‌های مثبت، خنثی و منفی جالب توجه است. برخی اصطلاحات منفی برجسته عبارتند از: "panic buying" و "toilet paper". در اصطلاحات مثبت، "hand sanitizer" توجه را به خود جلب می‌کند.

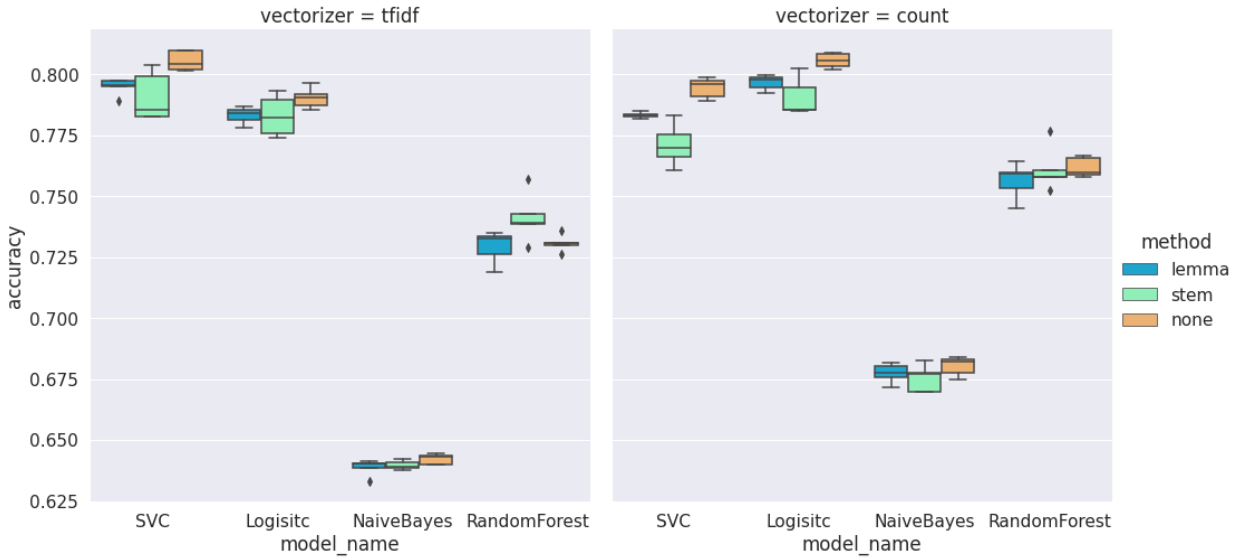
حال برای سهولت کار، توییت‌هایی را که Extremely Negative هستند به Negative و آن‌هایی را که Extremely Positive هستند به Positive تغییر می‌دهیم.

پیش از آن که به سراغ یادگیری عمیق برویم، ابتدا عملکرد چهار دسته‌بند مختلف را بررسی می‌کنیم که عبارتند از: SVC، $Logistic\ Regression$، $Naive\ Bayes$ و $Random\ Forrest$. همچنین این موضوع را بررسی می‌کنیم که استفاده از TF/IDF عملکرد بهتری دارد یا استفاده از $Count\ Vector$.

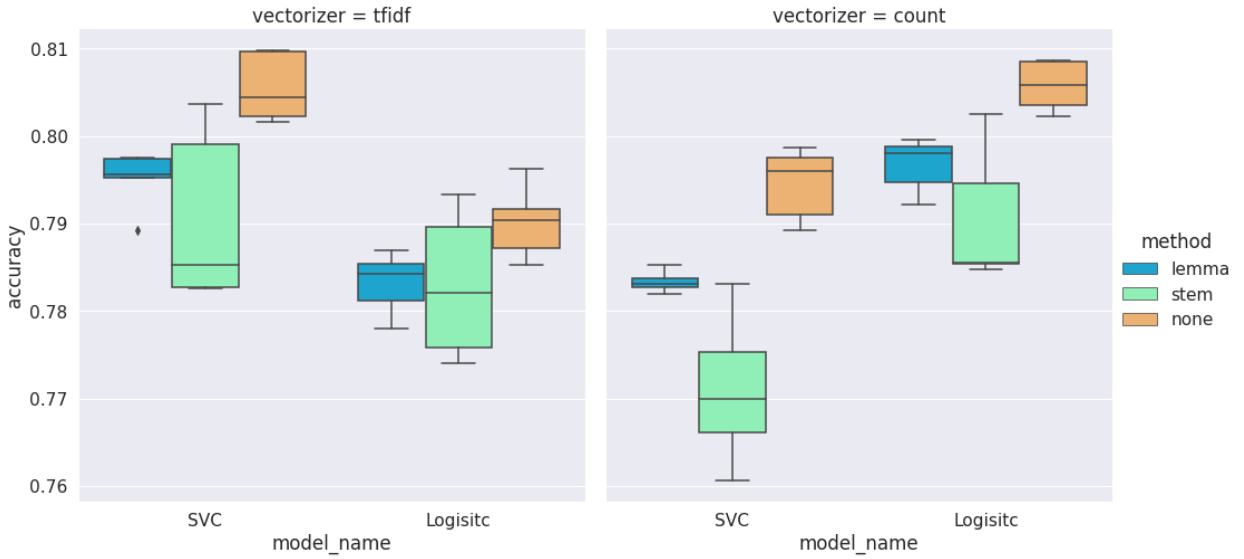
مقدار TF/IDF هر بار که یک کلمه در سند (توییت) دیده می‌شود، افزایش می‌یابد اما سپس این مقدار برای سندی که آن کلمه در آن دیده شده است، متعادل می‌شود. به این ترتیب، می‌توان کلماتی را انتخاب کرد که نقش کلیدی‌تری در دسته‌بندی داشته باشند. علاوه بر آن، با استفاده از Cross-Validation میزان دقت و واریانس هر مدل را روی چندین تقسیم داده، به دست می‌آوریم.

آن‌طور که مشاهده می‌شود، $Logistic\ Regression$ و $Random\ Forrest$ زمان بسیار بیشتری را نسبت به SVC و $Naive\ Bayes$ صرف می‌کنند.

	model_name	fold	accuracy	vectorizer	method
0	SVC	0	0.809781	tfidf	none
1	SVC	1	0.801675	tfidf	none
2	SVC	2	0.802216	tfidf	none
3	SVC	3	0.809646	tfidf	none
4	SVC	4	0.804377	tfidf	none
5	Logistic	0	0.791678	tfidf	none
6	Logistic	1	0.785328	tfidf	none
7	Logistic	2	0.790327	tfidf	none
8	Logistic	3	0.796271	tfidf	none
9	Logistic	4	0.787220	tfidf	none



همان‌طور که مشاهده می‌شود، $Naive\ Bayes$ و $Random\ Forrest$ در مقابل SVC و $Logistic\ Regression$ عملکرد بسیار ضعیفی دارند. به منظور این که بتوانیم نمودارهای جعبه‌ای[9] بهتر و با جزئیات بالاتری ببینیم، $Naive\ Bayes$ و $Random\ Forrest$ را حذف می‌کنیم.



می‌توان دید که SVC با استفاده از TF/IDF و $Logistic\ Regression$ با استفاده از $Count$ بهتر عمل می‌کند. گویا دقت stemming از lemmatization کمتر است اگرچه که lemmatization داده‌های خارج از محدوده[10]ی بیشتری دارد. اما بهترین نتیجه هنگامی به دست می‌آید که از هیچ‌کدام از stemming و lemmatization روی توییت‌ها استفاده نشده باشد.

[9] Boxplots
[10] Outliers

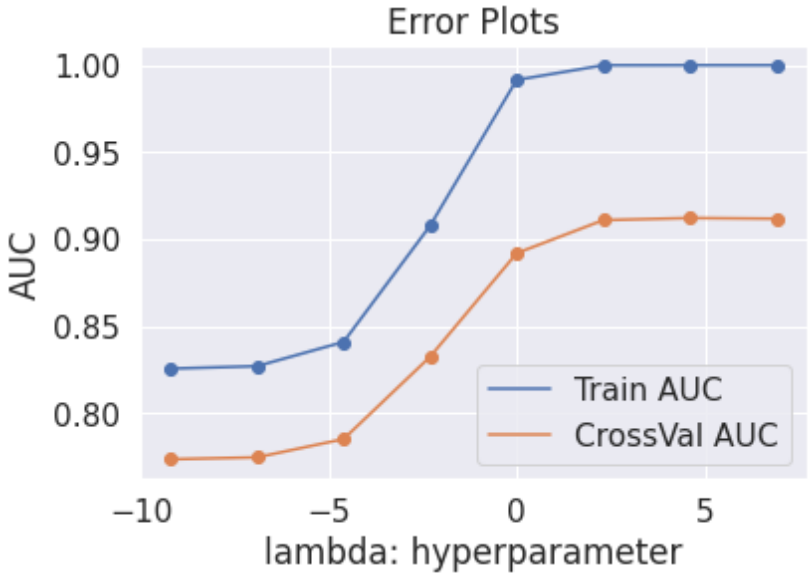
SVC با استفاده از TF/IDF و $Logistic\ Regression$ با استفاده از $Count$ تقریباً دارای میانه[11] یکسانی هستند اما SVC واریانس کمتری دارد و اندکی توزیع یکنواخت‌تری[12] دارد.

اختلاف دقت میان مدل‌هایی که بهترین عملکرد را دارند، بسیار کم است و احتمالاً بیشتر به دلیل تصادفی بودن تقسیم داده‌ها است و به روش و مدل ربطی ندارد.

با تمامی این اوصاف، استفاده از SVC به همراه TF/IDF بدون lemmatization و stemming به استفاده از $Logistic\ Regression$ ترجیح داده می‌شود؛ زیرا SVC زمان اجرای بسیار کمتری از $Logistic\ Regression$ دارد. علاوه بر آن، با توجه به نتایج حاصل شده، واریانس کمتری نیز دارد.

model_name	method	vectorizer	mean_acc	mean_std
Logistic	**none**	**count**	0.805782	0.002911
SVC	**none**	**tfidf**	0.805539	0.003943
Logistic	**lemma**	**count**	0.796644	0.003093
SVC	**lemma**	**tfidf**	0.795003	0.003414
	none	**count**	0.794488	0.004144
	stem	**tfidf**	0.790626	0.009970
Logistic	**stem**	**count**	0.790596	0.007830
	none	**tfidf**	0.790165	0.004233
SVC	**lemma**	**count**	0.783330	0.001246
Logistic	**lemma**	**tfidf**	0.783117	0.003587

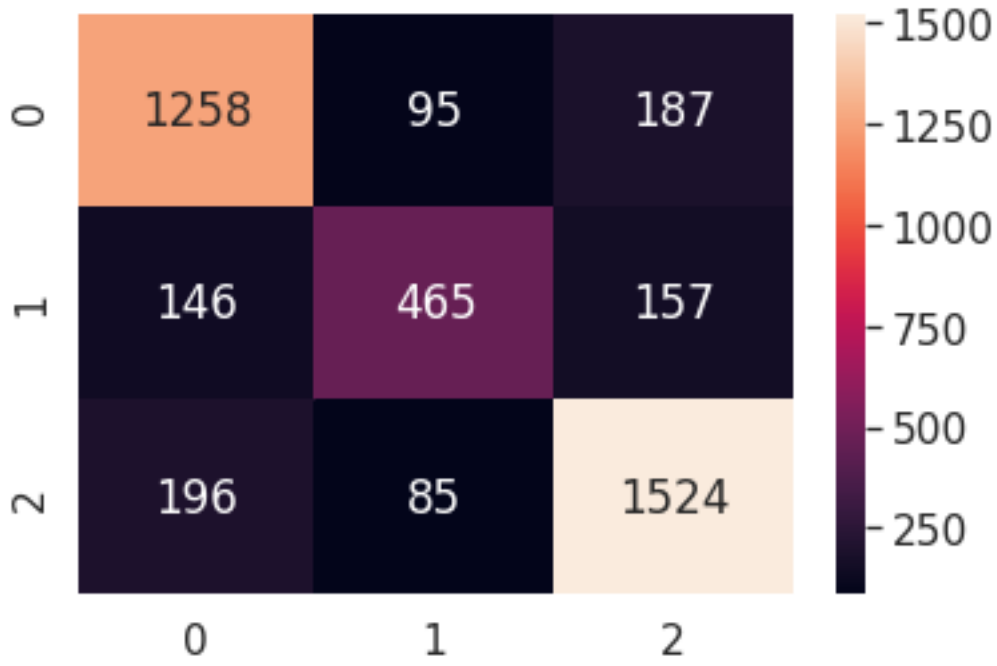
پارامترهای بهینه‌ی $Logistic\ Regression$ و نمودار AUC:



[11] Median
[12] Even Distribution

```
optimal_inverse_lambda: 0.01
AUC for Train set:     0.9999952984198565
AUC for Test set:      0.9137323441430732


Confusion Matrix of Test Data
[[1258   95  187]
 [ 146  465  157]
 [ 196   85 1524]]
Accuracy Score on test: 0.7894480914174569
```



	precision	recall	f1-score	support
Negative	0.79	0.82	0.80	1540
Neutral	0.72	0.61	0.66	768
Positive	0.82	0.84	0.83	1805
accuracy			0.79	4113
macro avg	0.77	0.76	0.76	4113
weighted avg	0.79	0.79	0.79	4113

حال شبکه‌ی عصبی از نوع $LSTM$ را بر روی داده‌های آموزش می‌دهیم. مشاهده می‌شود که پس از پنجمین epoch تغییرات در دقت شبکه‌ی عصبی بسیار اندک است و از آن صرف نظر می‌شود.

```
Epoch 1/10
405/405 [==============================] - 336s 821ms/step - loss: 0.9003 -
accuracy: 0.5586 - val_loss: 0.5453 - val_accuracy: 0.7881
Epoch 2/10
405/405 [==============================] - 327s 807ms/step - loss: 0.3801 -
accuracy: 0.8712 - val_loss: 0.4928 - val_accuracy: 0.8260
Epoch 3/10
405/405 [==============================] - 328s 810ms/step - loss: 0.2203 -
accuracy: 0.9304 - val_loss: 0.5518 - val_accuracy: 0.8152
Epoch 4/10
405/405 [==============================] - 329s 813ms/step - loss: 0.1577 -
accuracy: 0.9509 - val_loss: 0.6175 - val_accuracy: 0.8131
Epoch 5/10
405/405 [==============================] - 328s 810ms/step - loss: 0.1099 -
accuracy: 0.9670 - val_loss: 0.6898 - val_accuracy: 0.7968
```

سپس عملکرد شبکه‌ی عصبی را با استفاده از داده‌های آزمون[13] می‌سنجیم و مشاهده می‌شود که دقت شبکه‌ی عصبی بیش از ۸۰٪ است:

```
386/386 [==============================] - 29s 74ms/step - loss: 0.7030 -
accuracy: 0.8020
Test set
        Loss: 0.703
        Accuracy: 0.802
```

کارهای آینده

در آینده می‌توان درباره‌ی توییت‌هایی که دارای برچسب اشتباه[14] هستند، بیشتر تحقیق کرد و به مطالعه‌ی دقیق‌تر آن‌ها پرداخت. می‌توان تأثیر کاراکترهای خاص، مانند ‘Â’، در نظر گرفت تا خطای مدل کاهش یابد.

می‌توان با بررسی حروف بزرگ و در نظر گرفتن آن‌ها، مانند فرق کلمه‌ی much با MUCH، دقت مدل را افزایش داد.

در پایان می‌توان برای افزایش قابل توجه دقت مدل، از مدل $BERT$ استفاده کرد.

منابع

Koyel Chakraborty, S. B. (2020, December). Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied Soft Computing*. doi:https://doi.org/10.1016/j.asoc.2020.106754

[13] Test
[14] Mislabel