دانشگاه شهید بهشتی

دانشکده علوم پایه

گزارش پروژه درس نظریه یادگیری ماشین

تحلیل احساسات توییت های شرکت های هواپیمایی

Sentiment Analysis of Airline Twitter data

دانشجو: زهرا اروجی

استاد: دکتر حسین حاجی ابوالحسن

مقطع: کارشناسی ارشد

گرایش: علوم داده ها

تیرماه ۱۳۹۹

فهرست عناوین **صفحه**

۱ مقدمه و طرح مسأله

در دنیای امروز، اینترنت حجم وسیعی از اطلاعات را در بر گرفته است. افراد اینترنت را بـه عنـوان منبـع مهمی درک کرده اند که در آن تعداد زیادی از نظرات و تجربیات به آسانی در دسترس هستند. ارزیـابی های مردم به طور قابل توجهی بر باورها، برداشت‌ها و به ویـژه تصـمیمات خریـد آنهـا تـأثیر مـی‌گـذارد. امروزه، جریان و گردش اطلاعات به تدریج، به تجمع آنلاین تجربیات، بینش‌ها و دیدگاه‌هـا تبـدیل شـده است. افزایش شدید اطلاعات آنلاین فرصت قابل توجهی را برای شرکت‌ها ایجاد مـی‌کنـد تـا بهتـر درک کنند که مشتریان درباره یک محصول، موضوع یا نهاد دیگر چه می‌گویند.

در سال‌های اخیر، تجزیه و تحلیل احساسات توییتر برای تجزیـه و تحلیـل خودکـار رضـایت مشـتری از خدمات آنلاین بسیار رایج شده است. بازخورد مشتری در مورد خدمات آنها، بـرای شـرکت‌هـا مخصوصـاً شرکت‌های هواپیمایی بسیار ضروری است.

در این پروژه، داده‌های توییتر چند شرکت هواپیمایی مورد تجزیه و تحلیل قرار گرفته است کـه عـواملی چون بهترین و بدترین خطوط هواپیمایی، بیشترین دلایل نارضایتی مشـتریان، نظـرات مثبـت، منفـی و خنثی آنها درباره‌ی خدمات شرکت های هواپیمایی، در محـیط google colab بـا اسـتفاده از الگـوریتم های یادگیری ماشین، پیش پردازش، بصری‌سازی ،ارزیابی و توسط شبکه عصبی مدل و پیش‌بینـی مـی شود.

۲ تاریخچه

با رشد سریع شبکه های اجتماعی و بحث و گفتگوی آنلاین ، وب غنی از داده های متن آزاد تولید شـده توسط کاربر است، جایی که کاربران می توانند نگرش های مختلفی را نسبت به محصولات ابراز کنند. که این موضوع باعث شده است محققان به سمت تحلیل احساسات جذب شوند.

برای تعیین اینکه یک متن یا یک جمله بیانگر احساسات مثبت یـا منفـی باشـد ، معمـولاً از دو رویکـرد اصلی استفاده می شود: رویکرد مبتنی بر واژگان و رویکرد مبتنی بر یادگیری ماشین. رویکرد مبتنـی بـر واژگان شامل محاسبه جهت گیری برای یک متن از جهت معنای کلمات یا عبارات موجود در متن است. رویکرد مبتنی بر یادگیری ماشین شامل ایجاد classifiers از نمونه هـای دارای برچسـب مـتن هـا یـا

جملات است.اغلب این تکنیک‌ها مانند Maximum Entropy, Naive Bayes و Sopport Vector Machines در زمینه supervised learning هستند.

بطور مثال Sreenivasan و همکاران روی توییت‌های ۳ شرکت هواپیمایی مطالعه کردند. آنها از توییتر به عنوان منبع داده برای تحلیل ارتباطات مصرف کنندگان در مورد خدمات هواپیمایی استفاده کردند.

همینطور Breen و همکاران طبقه‌بندی احساسات توییت‌ها را با استفاده از واژه‌نامه احساساتی نمایش دادند و پیشنهاد دادند توییت‌های دارای زمان واقعی از Twitter API به جای پرسش های حاوی نمایش نام شرکت‌های هواپیمایی بازیابی شوند. در این روش آنها به دقت ۸۶/۴ ٪ رسیدند. و همچنین Adeborna و همکاران در رویکرد Naive Bayesian ،دو روش SVM وEntropy را مقایسه کرده اند.

۳ شرح داده

مجموعه داده مورد نظر ما که از سایت آموزشی kaggle دریافت شده است به فرمت csv می‌باشد که شامل ۱۴۶۴۰ توییت از ۷۷۰۰ کاربر و مجموعاً ۱۵ ویژگی است. داده‌های توییتر از فوریه ۲۰۱۵ حذف شد و از مشتریان خواسته شد ابتدا توییت‌های مثبت ، منفی و خنثی را ثبت کنند و به دنبال آن دسته بندی دلایل منفی (مانند "پرواز دیررس" یا "سرویس دهی بد") را اشتراک گذاری کنند.

۴ شرح روش

در ابتدا کتابخانه های مورد نیاز و اساسی همچون scikit learn و keras و همچنین مجموعه داده را فراخوانی می‌کنیم. سپس به پردازش و پاک سازی داده و استخراج ویژگی های کاربردی یا استفاده از تحلیل نمودارها و درک مجموعه داده‌ها می‌پردازیم.سپس با استفاده از روش امتیازدهی به نظرات مثبت و منفی، مجموعه داده‌ها را برای آموزش آماده کرده و در نهایت الگوریتم منتخب یادگیری ماشین را روی مجموعه داده‌های حاصل به‌کار می‌گیریم. در نهایت به ارزیابی خروجی مدل می‌پردازیم.

۴-۱ پیش پردازش

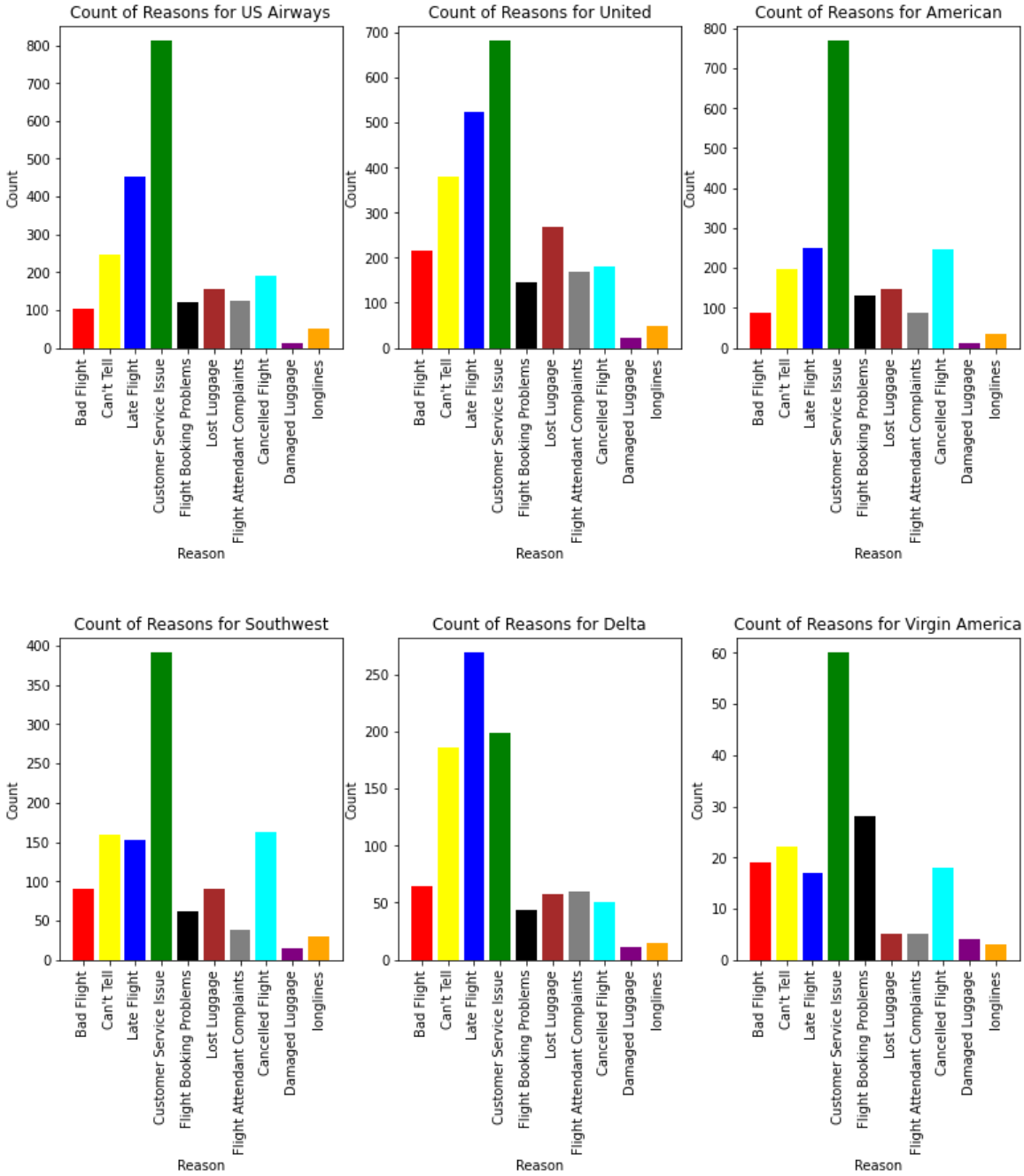
الف- مجموعه داده ما شامل ۱۵ ویژگی (ستون) است که جزئیات آن به شرح زیر است:

feature	Details		
tweet_id	14640	non-null	int64
airline_sentiment	14640	non-null	object
airline_sentiment_confidence	14640	non-null	float64
negativereason	9178	non-null	object
negativereason_confidence	10522	non-null	float64
airline	14640	non-null	object
airline_sentiment_gold	40	non-null	object
name	14640	non-null	object
negativereason_gold	32	non-null	object
retweet_count	14640	non-null	int64
text	14640	non-null	object
tweet_coord	1019	non-null	object
tweet_created	14640	non-null	object
tweet_location	9907	non-null	object
user_timezone	9820	non-null	object

ب- پس از بررسی مقادیر گمشده ویژگی ها، به این نتیجه می‌رسیم که ۳ ویژگی یعنی **airline_sentiment_gold** و **negativereason_gold** و **tweet_coord** دارای درصد زیادی داده گمشده هستند پس آنها را از میان ویژگی ها حذف می‌کنیم و بدون درنظرگرفتن آنها به پردازش ادامه می‌دهیم.

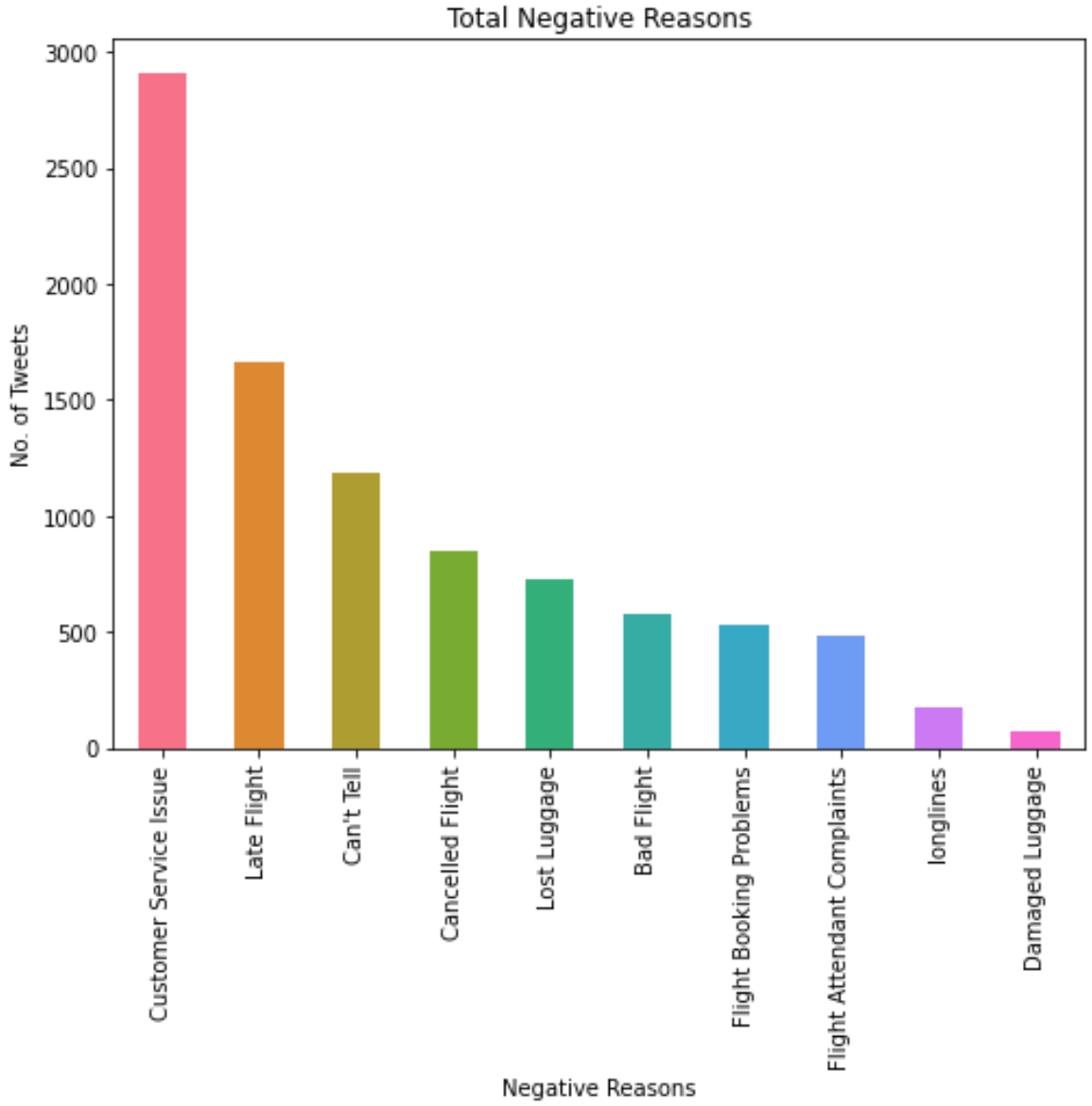
۴ – ۲ تحلیل بصری و نمودارها

برای بررسی دلایل منفی مشتریان برای همه شرکت‌های هواپیمایی نمودار میله‌ای آنها را رسم می‌کنیم و به مقایسه آنها می‌پردازیم:

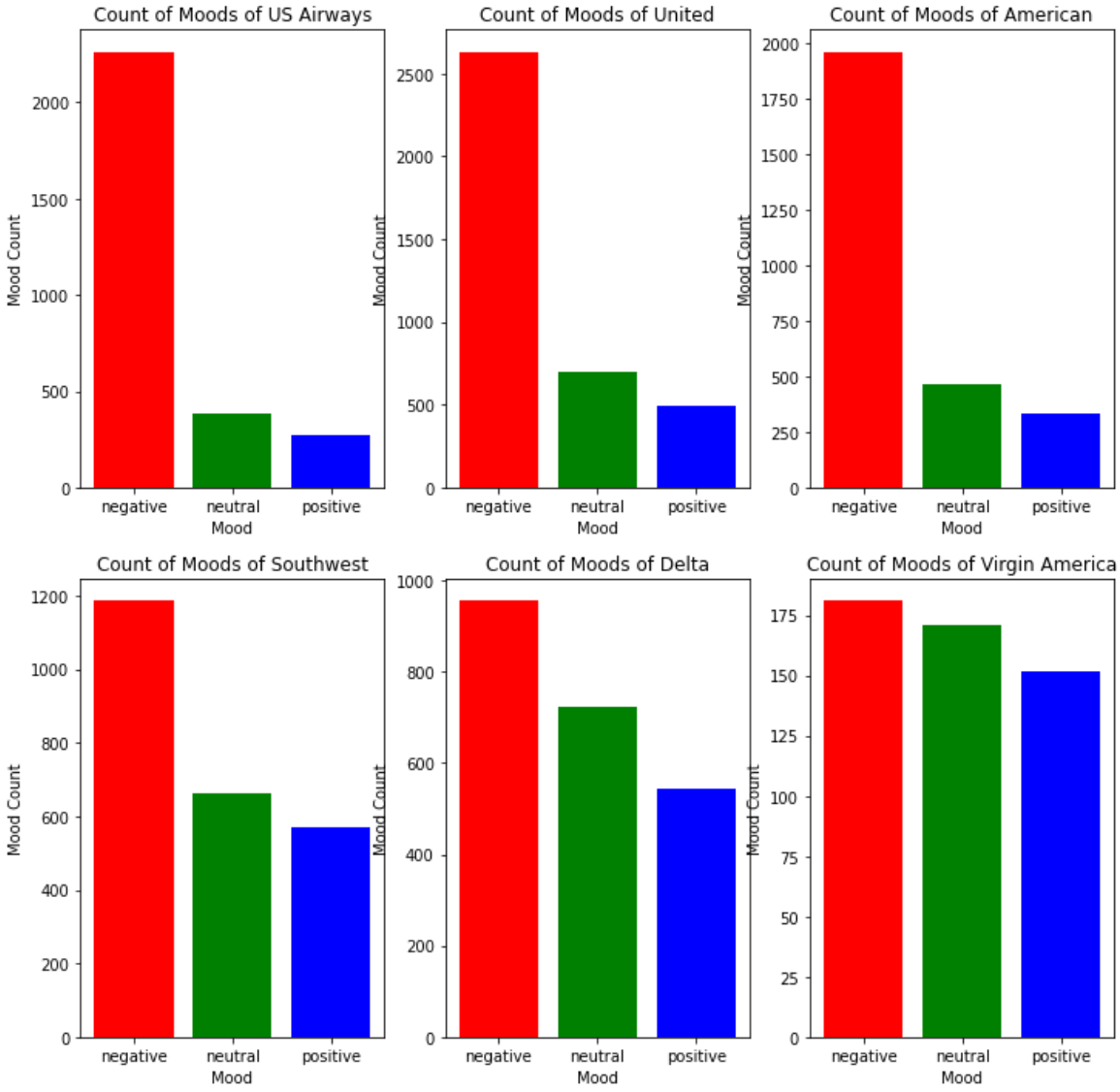


با مقایسه نمودار ها درمی‌یابیم عمده تـرین دلیـل بـرای توییـت منفـی مشـتریان در همـه شـرکت هـا، **Customer Service Issue**، است.

برای درک بهتر **تعداد نظرات منفی** نمودار زیر را به نمایش می‌گذاریم:

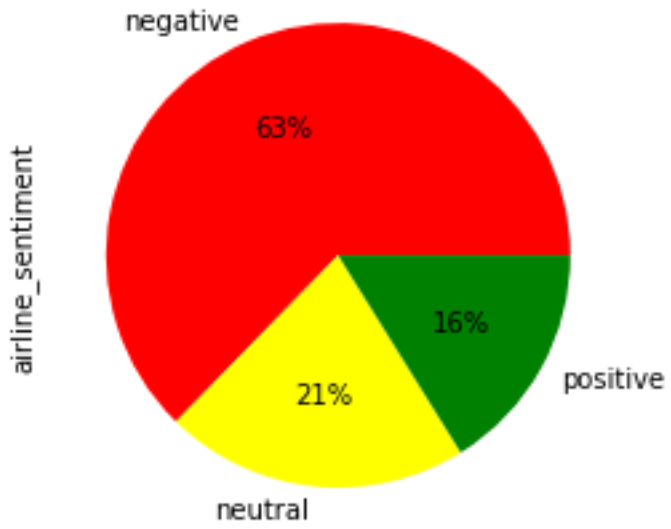


برای نمایش **احساسات مثبت، منفی و خنثی** مشتریان نسبت بـه خـدمات شـرکت هـای هواپیمـایی نمودار های میله ای آنها را بطور جداگانه رسم کردیم و به تحلیل آنها پرداختیم.



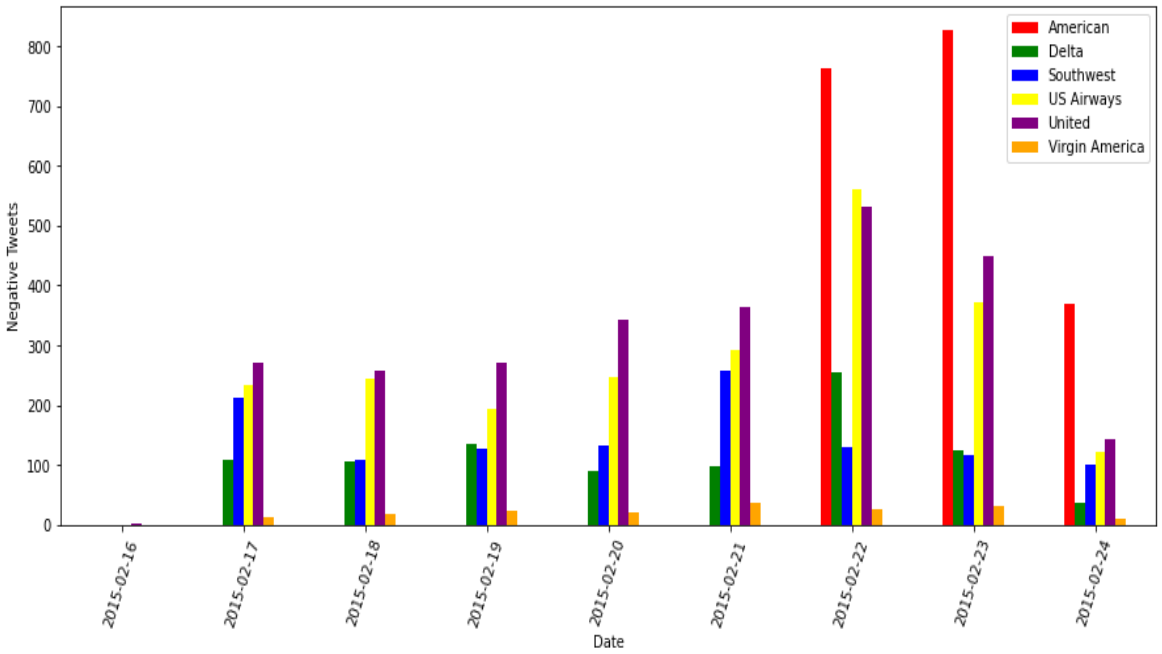
با تحلیل نمودارهای فوق دریافتیم شرکت‌های **United, US Airways, American** بطـور چشـم گیـری **واکنش منفی** دریافت می‌کنند. ولی **Virgin America** دارای واکنش متعادل‌تری است.

برای دستیابی به میزان دقیق واکنش های مشتریان **نمودار دایره ای** آن را رسم کردیم:



با توجه به نمودار فوق درمی‌یابیم مجموع ۶ شرکت هواپیمایی **۶۳٪ نظرات منفی، ۱۶٪ نظرات مثبت** و **۲۱٪ نظرات بی تفاوت و خنثی** را از مشتریان دریافت کردند.

همچنین **میزان دریافت نظرات و احساسات منفی مشتریان** را برحسب **تاریخ ثبت** آنها نیز بررسی می‌کنیم:



با بررسی تاریخ ثبت شده برای توییت‌های منفی در می‌یابیم شرکت هواپیمایی **American** بطور ناگهانی در تاریخ های **2015-02-22 و 2015-02-23** نظرات منفی دریافت کرده است.

همینطور **Virgin America** در مقایسه با سایر شرکت های هواپیمایی دارای حداقل نظرات منفی در طول هفته است.

و همه شرکت های هواپیمایی در آخر هفته دارای تعداد نظرات منفی بیشتری هستند.

۴ – ۳ جدا سازی دو ستون یا ویژگی برای اجرای مدل و پاک سازی آنها

ستون های text و airline_sentiment را برای بررسی داده های آموزش و تست جدا می‌کنیم و به کمک تابع های x.lower() و re.sub() و Tokenizer() به پاک سازی و ویرایش و آماده سازی آنها می‌پردازیم.

۴- ۴ تعیین داده های آموزش و داده های تست

داده های ویژگی airline_sentiment را به عنوان داده تست و text را به عنوان مجموعه آموزشی درنظر می‌گیریم.

۵ پیاده‌سازی مدل

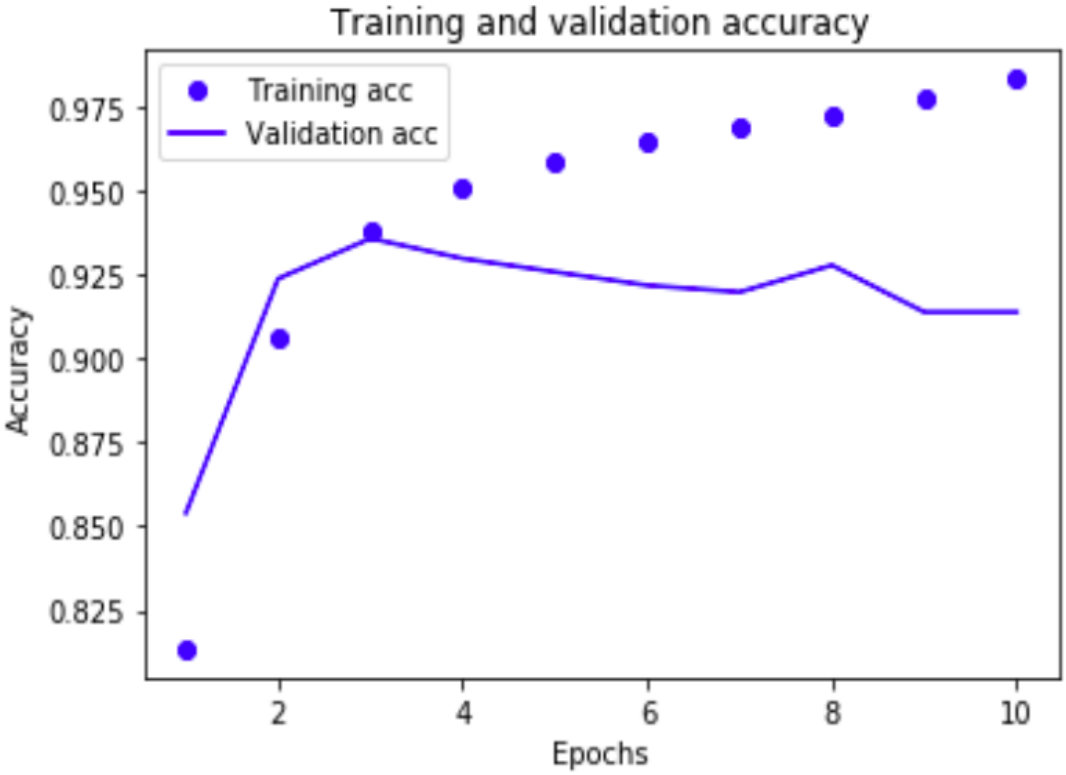
شبکه عصبی خود را با اطلاعات زیر طراحی و روی داده ها پیاده‌سازی می‌کنیم:

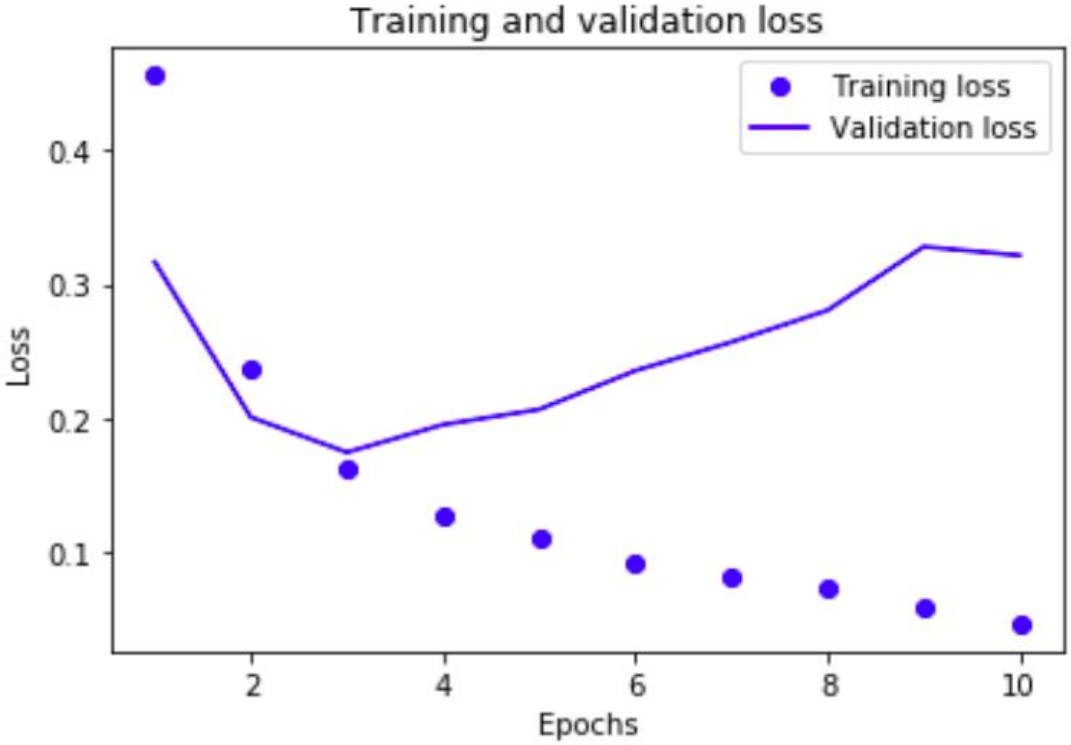
max_fatures = 2000
embed_dim = 128
lstm_out = 196
loss = categorical_crossentropy
optimizer=adam
metrics = accuracy
Dense= 2
activation=softmax
batch_size=512
epochs=10

۶ نتیجه‌گیری

با توجه به نمودار ها و نتایج مشاهده در مدل به این نتیجه پی می‌بریم که عملکرد مدل ما دارای میزان دقت ۹۴٪ برای نظرات منفی و ۸۲٪ برای نظرات مثبت است که این میزان دقت در مقایسه با سایر روش ها و مدل های دیگر یادگیری ماشین که توسط بنده و سایر محققان بررسی شده دارای میزان بسیار خوبی است.

نمودار نتایج:





منابع و مراجع

[۱] Haji H. BINALI, Chen WU, Vidyasagar POTDARA, “New Significant Area: Emotion Detection in E-learning Using Opinion Mining Techniques”, 3rd IEEE International Conference on Digital Ecosystems and Technologies, pp. 259-264, 2009

[۲] Bing Liu. Sentiment analysis and subjectivity. In Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca, 2010.

[۳] Hsu CW, L.C., A Simple Decomposition Method for Support Vector Machines. Machine Learning, 46, 291–314. URL http://www.csie.ntu.edu.tw/~cjlin/papers/decomp.ps.gz., 2002

[۴] Nan, L.a.D., D., Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems archive. Volume 48 Issue 2, January, 2010. Pages 354-368. Elsevier Science Publishers B. V. Amsterdam, The Netherlands, The Netherlands, 2010.

[۵] Sreenivasan, Nirupama Dharmavaram, Chei Sian Lee, and Dion Hoe-Lian Goh. "Tweeting the friendly skies: Investigating information exchange among Twitter users about airlines." Program: electroni.

[۶] Breen, Jeffrey Oliver. "Mining twitters for airline consumer sentiment." Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications , pp. 133, 2012

[۷] Adeborna, Esi, and Keng Siau. "An Approach to Sentiment Analysis–The Case of Airline Quality Rating." PACIS 2014 Proceedings, pp.363.