گزارش پروژه اول درس شناسایی الگو

برای انجام این پروژه، تمامی مراحل لازم و کار های انجام شده را در سلول های بالای کد ها توضیح داده شده است. توجه شود که برای کلاسیفایر ها و متریک ها از ماژول sklearn استفاده شده است. همچنین برای بدست آوردن بردار های BOW و TF_IDF نیز از ماژول sklearn استفاده کردیم. در دو روش BOW و TF_IDF 10000 تا ویژگی در نظر گرفته ایم.(زیرا برای مقادیر بیشتر مثلا 20000 تا با مشکل کمبود رم در colab مواجه شدیم.). در روش bert هر review را به جمله هایش tokenize کردیم و هر جمله را به مدل bert دادیم. سپس میانگین بردار های جمله های هر review را به عنوان بردار آن review در نظر گرفتیم.

با توجه به طولانی بودن اجرای روش svm برای هیچ یک از حالت ها موفق به اجرای این کلاسیفایر نشدیم.

نتیاج به صورت زیر است: (max هر ستون پر رنگ شده است.)

BOW	AUC	ROC	Precision	Recall	F1_score
Naïve Bayes	**0.7707891684332626**	(array([0. , 0.07959682, 1.]), array([0. , 0.62117515, 1.]), array([2., 1., 0.]))	**0.8864155251141552**	0.621175152993 8803	0.730462594953082
SVM					
Decision tree	0.7188912443 50226	(array([0. , 0.274949, 1.]), array([0. , 0.71273149, 1.]), array([2., 1., 0.]))	0.7216215121694407	**0.7127314907403703**	0.7171489515836922
Random Forest	0.7632894684 21263	(array([0. , 0.17531299, 1.]), array([0. , 0.70189192, 1.]), array([2., 1., 0.]))	0.8001459121791072	0.7018919243230 27	**0.7478053353788461**

bert	AUC	ROC	Precision	Recall	F1_score
Naïve Bayes	**0.8359065637374505**	(array([0. , 0.13987441, 1.]), array([0. , 0.81168753, 1.]), array([2. , 1., 0.]))	**0.8530054644808743**	**0.8116875324987001**	**0.8318337398290668**
SVM					
Decision tree	0.7783088676452942	(array([0. , 0.21811128, 1.]), array([0. , 0.77472901, 1.]), array([2. , 1., 0.]))	0.7803158488437676	0.7747290108395665	0.7775123938743153
Random Forest	0.8324467021319149	(array([0. , 0.14475421, 1.]), array([0. , 0.80964761, 1.]), array([2. , 1., 0.]))	0.8483299107329952	0.8096476140954362	0.8285375138144161

TF_IDF(Removing stop words)	AUC	ROC	Precision	Recall	F1_score
Naïve Bayes	0.7162513499460021	(array([0. , 0.18175273, 1.]), array([0. , 0.61425543, 1.]), array([2., 1., 0.]))	0.7716697653384252	0.6142554297828087	0.6840229833860407
SVM					
Decision tree	0.7057717691292349	(array([0. , 0.28838846, 1.]), array([0. , 0.699932, 1.]), array([2., 1., 0.]))	0.7082034886073 9	0.6999320027198912	0.7040434520217 26
Random Forest	**0.7805087796488 14**	(array([0. , 0.14387425, 1.]), array([0. , 0.7048918, 1.]), array([2., 1., 0.]))	**0.8304901036757776**	**0.7048918043278268**	**0.7625538175288289**

TF_IDF(without removing stop words)	AUC	ROC	Precision	Recall	F1_score
Naïve Bayes	0.71673133074677702	(array([0. , 0.18463261, 1.]), array([0. , 0.61809528, 1.]), array([2., 1., 0.]))	0.7699935223478998	0.6180952761889524	0.6857333037497226
SVM					
Decision tree	0.7026118955241789	(array([0. , 0.29338826, 1.]), array([0. , 0.69861206, 1.]), array([2., 1., 0.]))	0.7042457965404662	**0.6986120555177793**	0.7014176137504518
Random Forest	**0.7489900403983841**	(array([0. , 0.16683333, 1.]), array([0. , 0.66481341, 1.]), array([2., 1., 0.]))	**0.7993939976914197**	0.6648134074637014	**0.7259188085515255**

از لحاظ معیار AUC در روش BOW کلاسیفایر naïve bayes از همه بهتر بوده است.

از لحاظ معیار AUC در روش bert کلاسیفایر naïve bayes باز هم از همه بهتر بوده است.

از لحاظ معیار AUC در روش tf idf بدون حذف و با حذف کلمات توقفی کلاسیفایر Random forestاز همه بهتر بوده است.

در کل بهترین AUC را به ترتیب روش های bert و tf idf با حذف کلمات توقفی و BOW و tf idf بدون حذف کلمات توقفی است.

با مقیسه

از لحاظ معیار f1 score در روش BOW کلاسیفایر naïve bayes از همه بهتر بوده است.

از لحاظ معیار f1 score در روش bert کلاسیفایر naïve bayes باز هم از همه بهتر بوده است.

از لحاظ معیار f1 score در روش tf idf بدون حذف و با حذف کلمات توقفی کلاسیفایر Random forestاز همه بهتر بوده است.

در کل بهترین f1 score را به ترتیب روش های bert و tf idf با حذف کلمات توقفی و BOW و tf idf بدون حذف کلمات توقفی است.

در نهایت بهترین روش استفاده از bert و کلاسیفایر naïve bayes است.