پروژه پایانی

حسن حمیدی
حامد همتیان
آرش لگزیان

تشخیص اخبار تاثیرگذار

بررسی و اهمیت موضوع

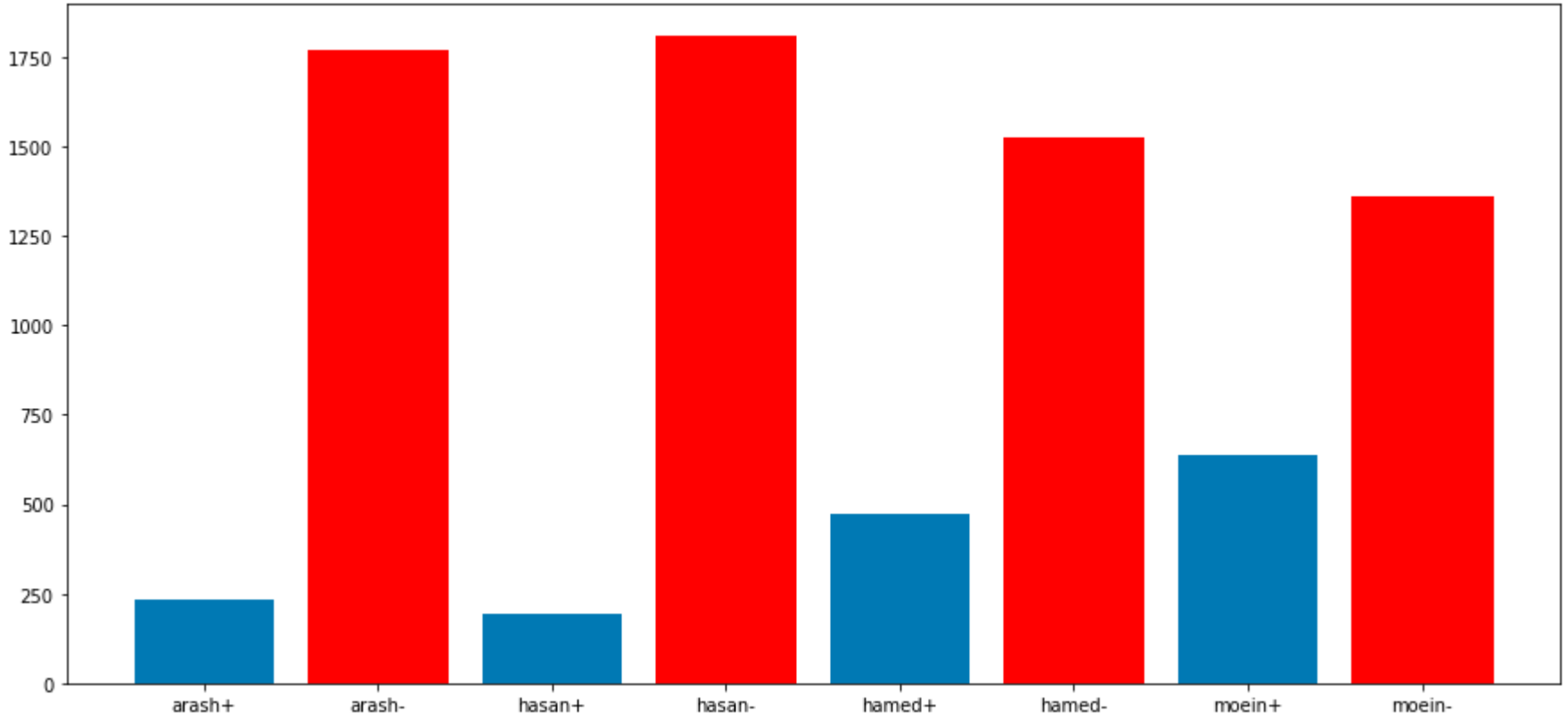
امروزه با وجود حجم بالای اخباری که در خبرگذاری های مختلف منتشر می‌شوند، افراد ترجیح می‌دهند فقط خبرهای مهم و تاثیرگذار را ابتدا بررسی کنند و سپس در صورت داشتن وقت کافی به مطالعه و دنبال کردن سایر خبرها بپردازند.

مرحله آماده‌سازی داده

- ما دو گروه, هر گروه شامل دو rater داشتیم.
- هر گروه ۲۰۰۰ خبر را برچسب زدند.
- سپس یک نفر سوم برای تگ زنی نهایی گروه دیگر استفاده شد.
- در مجموع ۴۰۰۰ خبر برچسب زده شد.

ساخت پیکره

توزیع برچسب‌های مثبت و منفی برای هر rater



ساخت پیکره

محاسبه معیار fleiss kappa برای هر category

class	agreement
اجتماعی	۰.۲۵
اقتصادی	۰.۱۱
بین الملل	۰.۱۰
حوادث	۰.۲۱
سیاسی	۰.۲۲
علمی و پزشکی	۰.۲۱
فرهنگ و هنر	۰.۱۹
فناوری و ارتباط	۰.۱۴
مذهبی	۰.۰۷
ورزشی	۰.۲۲

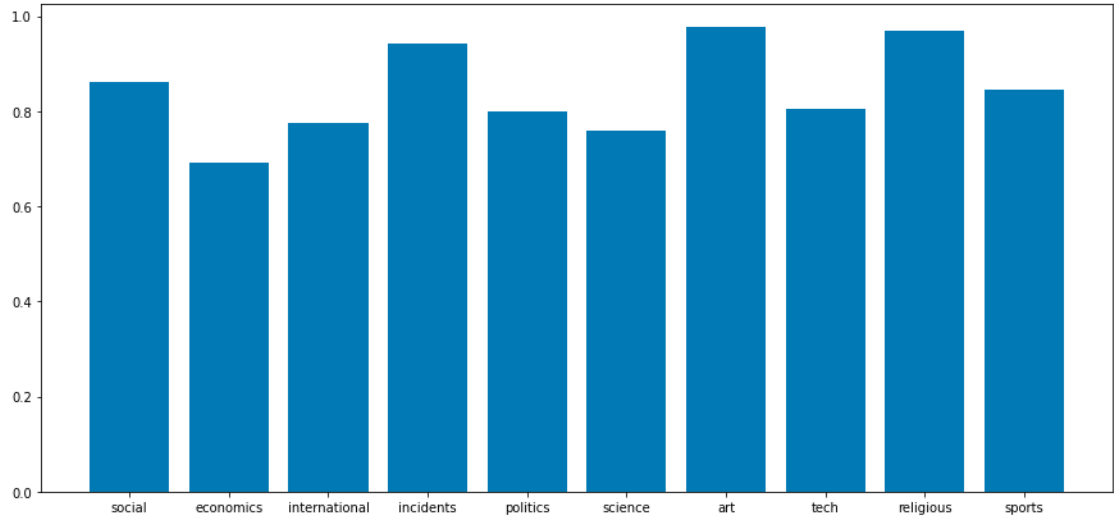
ساخت پیکره

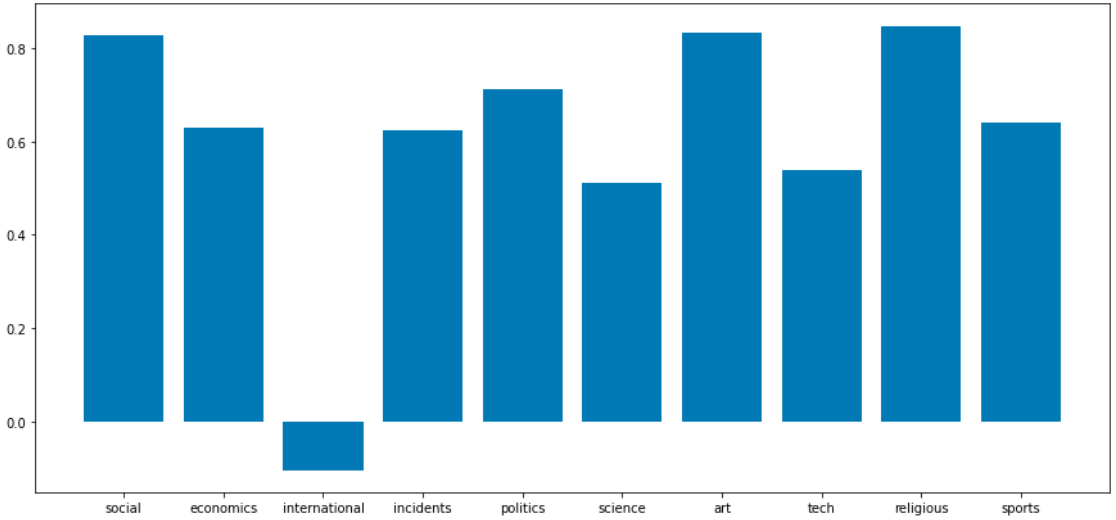
محاسبه معیار kappa برای هر category برای هرتیم

class	0.agreement	
	Hassan & Arash	Moein & Hamed
اجتماعی	0.86	0.82
اقتصادی	0.69	0.63
بین الملل	0.77	-0.10
حوادث	0.94	0.62
سیاسی	0.79	0.71
علمی و پزشکی	0.79	0.51
فرهنگ و هنر	0.97	0.83
فناوری و ارتباط	0.80	0.53
مذهبی	0.96	0.84
ورزشی	0.86	0.63
میانگین	0.83	0.64

ساخت پیکره

محاسبه معیار fleiss_kappa : ۰.۲۱
نمودار معیار agreement برای هر کلاس برای هرتیم





مرحله پیش پردازش و آماده‌سازی داده

- در این مرحله داده از دیاتفریم استخراج شده و هر داده به شکل یک شی از کلاس Data درمی‌آید.
- برچسب های غیرمهم و مهم نیز به 0 و 1 کدگذاری میشوند.
- برای آزمایش‌ها مجموعه داده به سه بخش آموزش و اعتبارسنجی و تست با نسبت‌های 0.6 و 0.2 و 0.2 تقسیم شد.
- برای تمامی مدل‌ها ما از 3 پیش‌پردلزش استفاده می‌کنیم که توسط کلاس Preprocessing اعمال میشود.
- این پیش‌پردازش شامل:
  - حذف punctuation
  - حذف اعداد
  - حذف stopwords
  - حذف فواصل اضافی

مدل‌های استفاده شده در این پروژه

✓ در این پروژه به بررسی ۷ مدل مختلف پرداخته شده است که در جدول زیر قابل مشاهده هستند.

class	model
Classic machine learning model	Logistic Regression
	Naive Bayes
	SVM
	Decision Tree
	Random Forest
Deep neural network model	BiLSTM-Attrntion based Net
Transformer	BERT
	Roberta

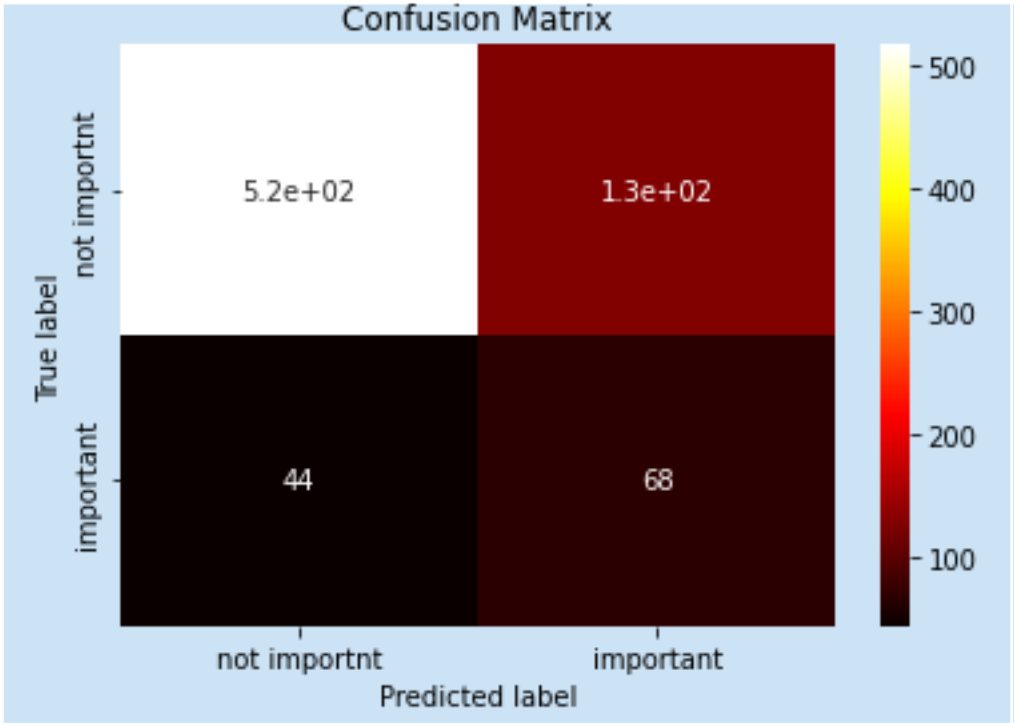
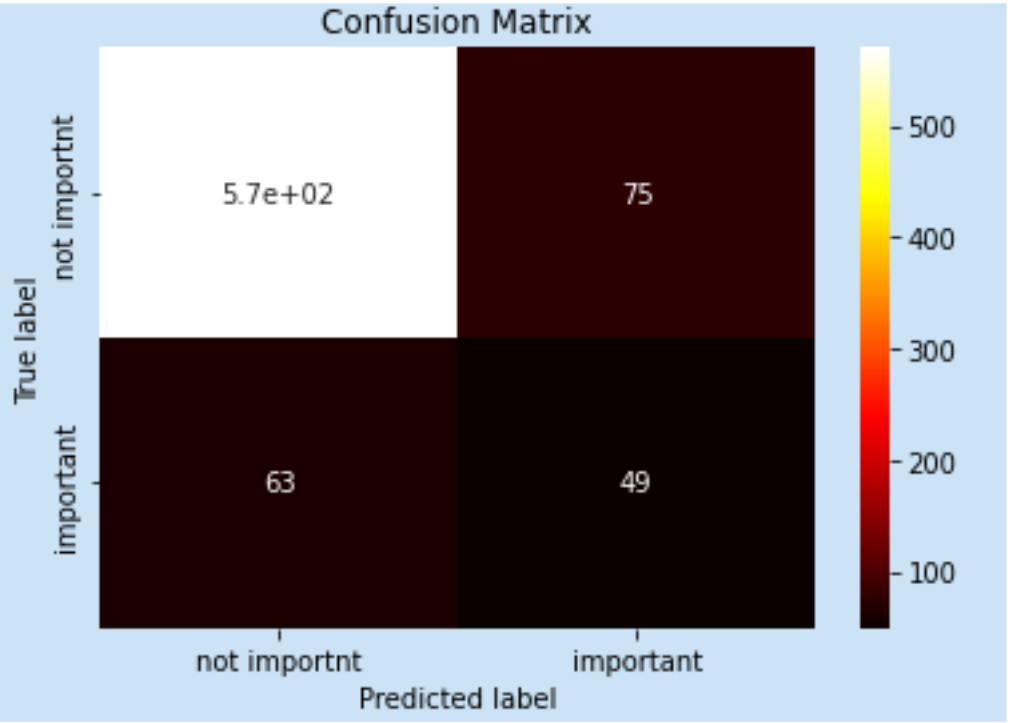
مدل Logistic Regression

- ✓ از cross-validation با c=5 جهت آموزش مدل و یافتن بهترین پارامترها استفاده شده است.
- ✓ زمان آموزش حدودا ۳۰ دقیقه

metrics	Before tuning		After tuning	
	TF-IDF	BoW	TF-IDF	BoW
F1-score macro	0.477	0.602	0.649	0.653
F1-score micro	0.853	0.845	0.772	0.817
Accuracy score	0.853	0.845	0.772	0.817
Recall score	0.508	0.584	0.704	0.660
Precision score	0.760	0.663	0.634	0.647
ROCAUC score	0.508	0.584	0.704	0.660
Grid Search parameters	-	-	C=1, class weight={0:0.15, 1:1}, Solver = newton-cg	C=0.01, class weight={0:0.15, 1:1}, Solver = newton-cg

مدل Logistic Regression

- ✓ ماتریس درهم ریختگی برای مدل fine tune شده.

TF-IDF	BoW
	

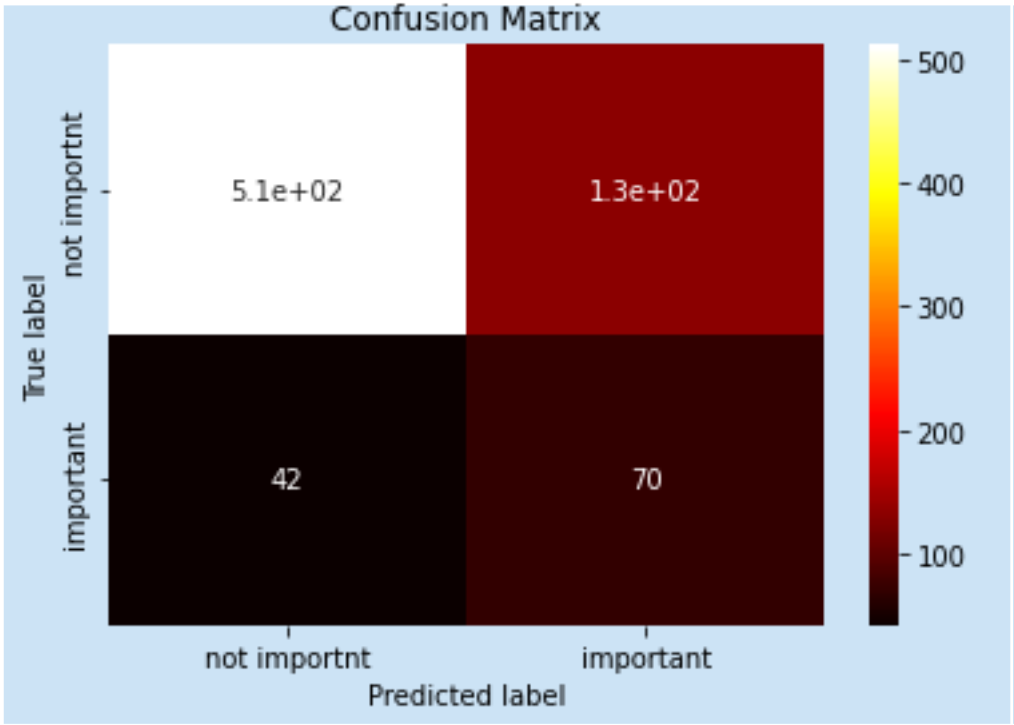
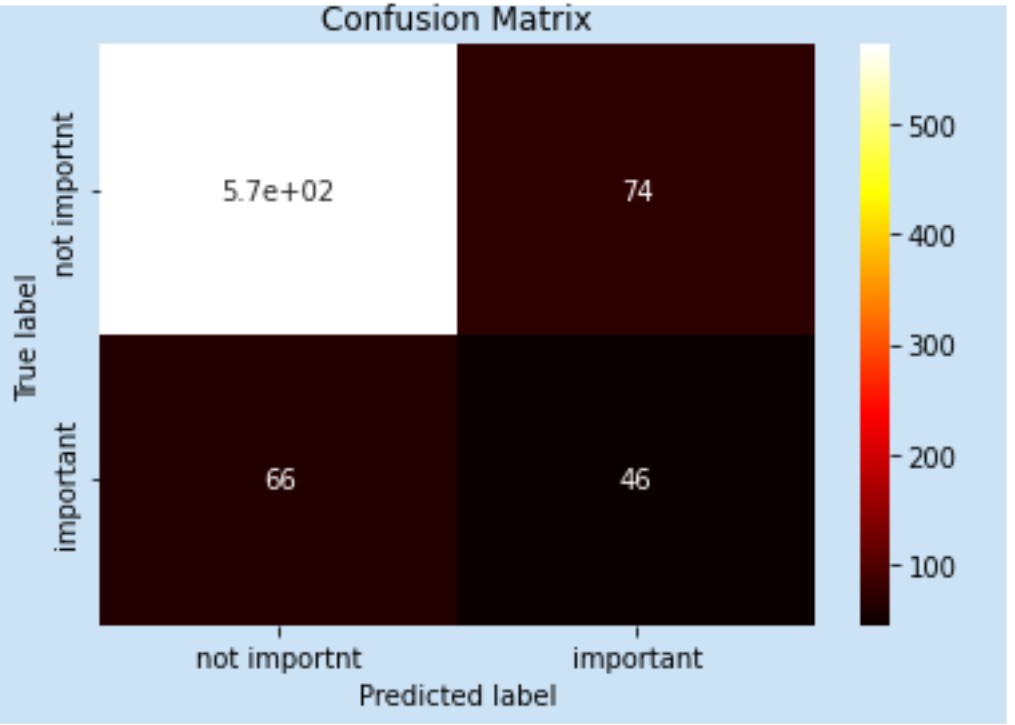
مدل SVM

- ✓ از cross-validation با c=5 جهت آموزش مدل و یافتن بهترین پارامترها استفاده شده است.
- ✓ زمان آموزش حدودا ۶.۵ ساعت.

metrics	Before tuning		After tuning	
	TF-IDF	BoW	TF-IDF	BoW
F1-score macro	0.486	0.460	0.649	0.643
F1-score micro	0.854	0.852	0.768	0.815
Accuracy score	0.854	0.852	0.768	0.815
Recall score	0.512	0.5	0.709	0.647
Precision score	0.802	0.426	0.634	0.639
ROCAUC score	0.512	0.5	0.709	0.647
Grid Search parameters	-	-	C=1, class weight={0:0.2, 1:3}, Degree=2, kernel=linear	C=0.01, class weight={0:0.15, 1:5}, Degree=2, kernel=linear

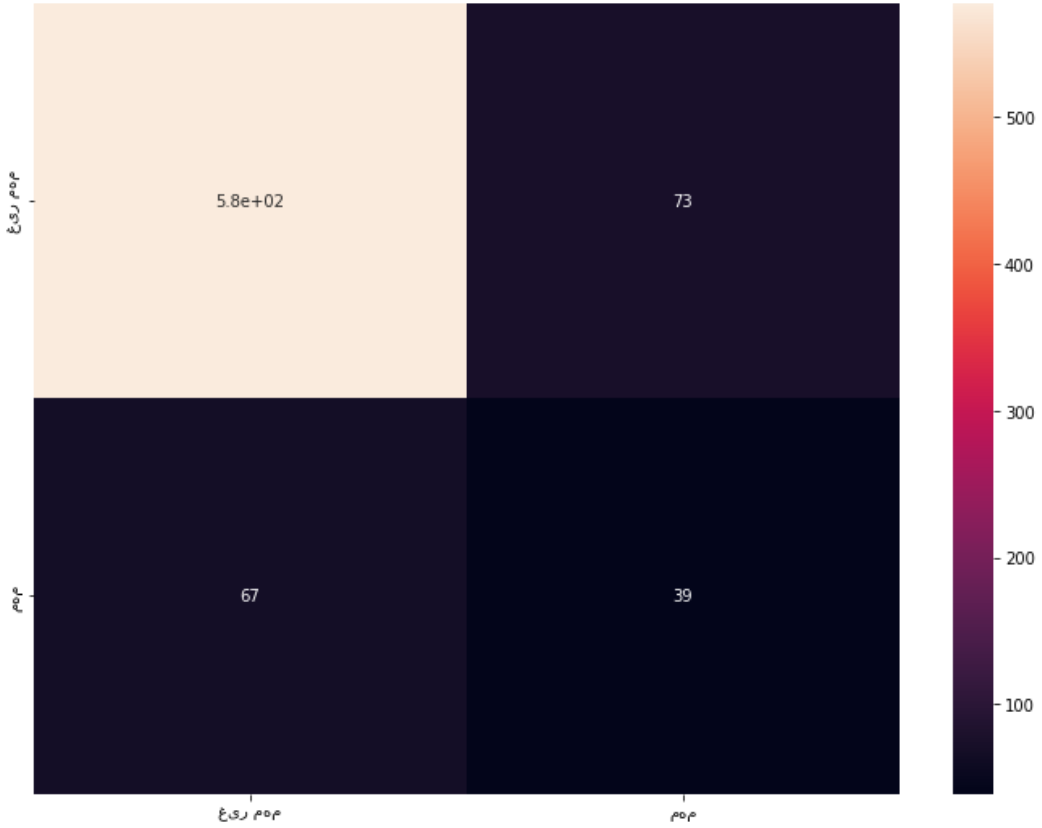
مدل SVM

- ✓ ماتریس درهم ریختگی برای مدل fine tune شده.

TF-IDF	BoW
	

نتایج مدل BiLSTM + Attention

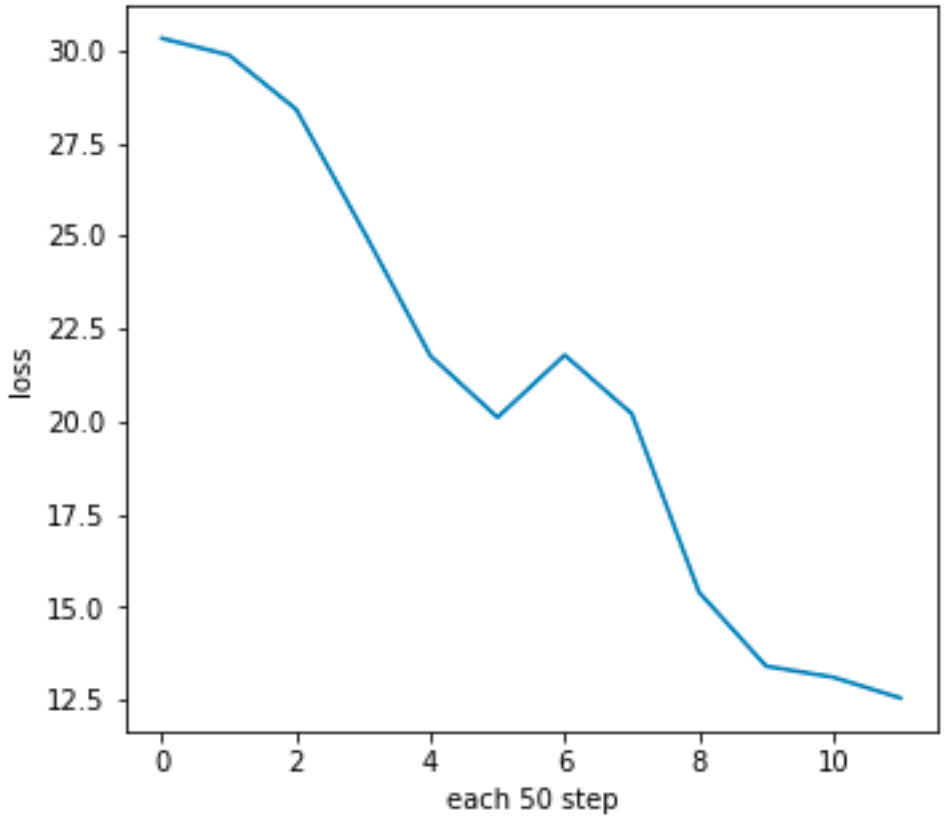
در این مدل یک bilstm به همراه دو مکانیزم توجه استفاده شده یک مکانیزم توجه برای لایه نهان جلو رونده یک مکانیزم توجه برای برای لایه نهان عقب رونده:

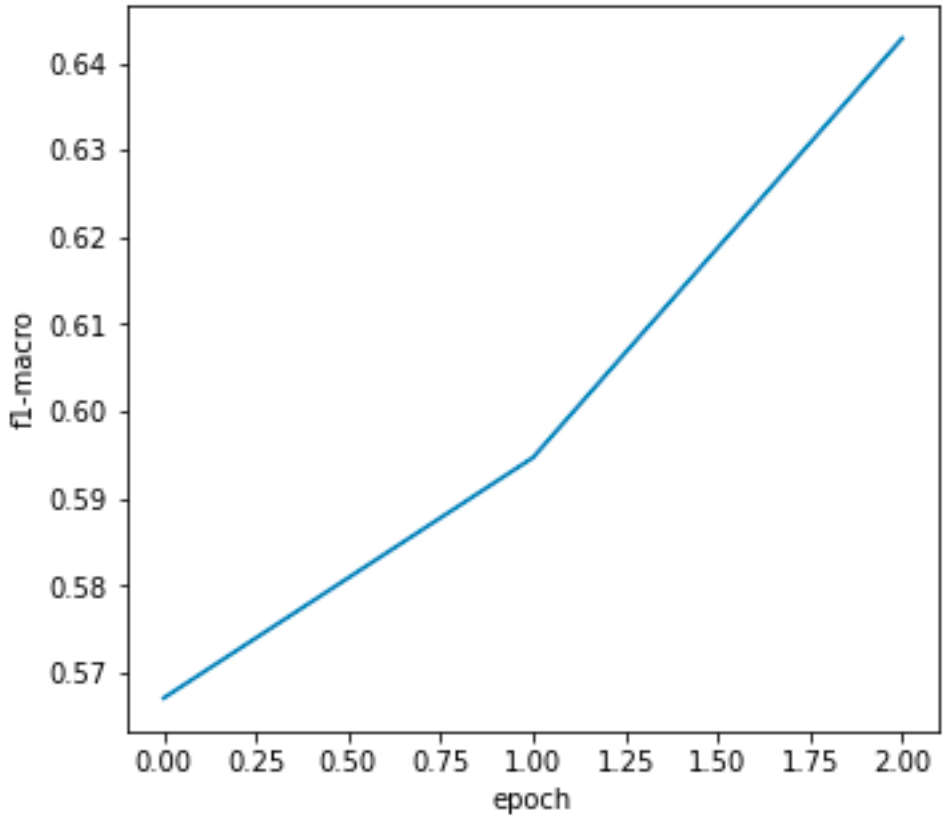


metrics	Test
F1-score macro	0.624
F1-score micro	0.815
Accuracy score	0.815
Recall score	0.348
ROCAUC score	0.622

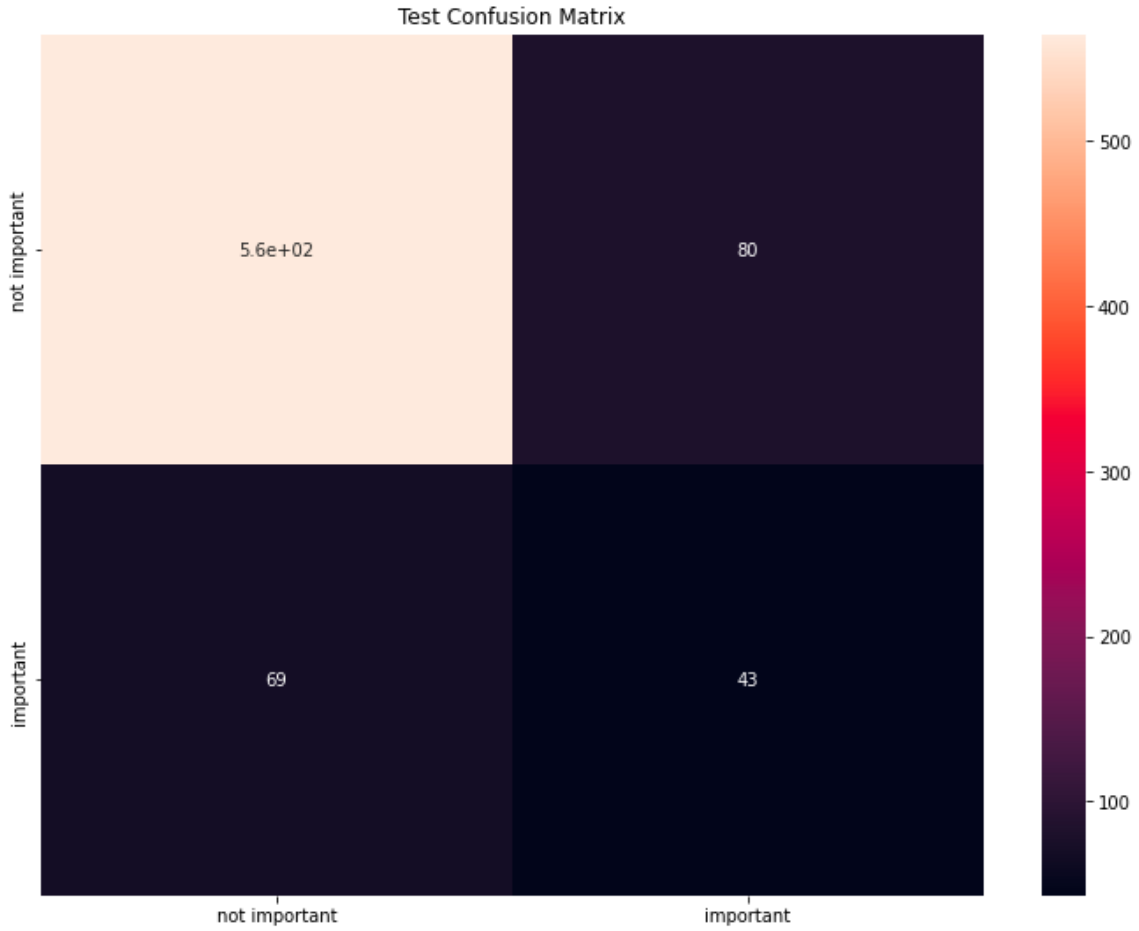
مدل Bert

نمودار هزینه و معبار ارزیابی f1-macro در زمان آموزش





مدل Bert

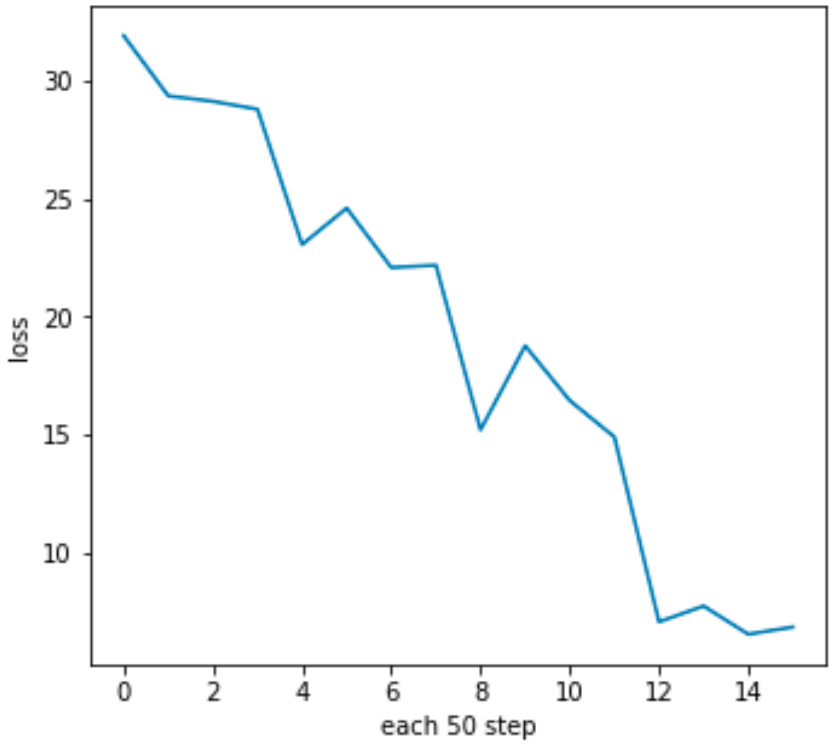


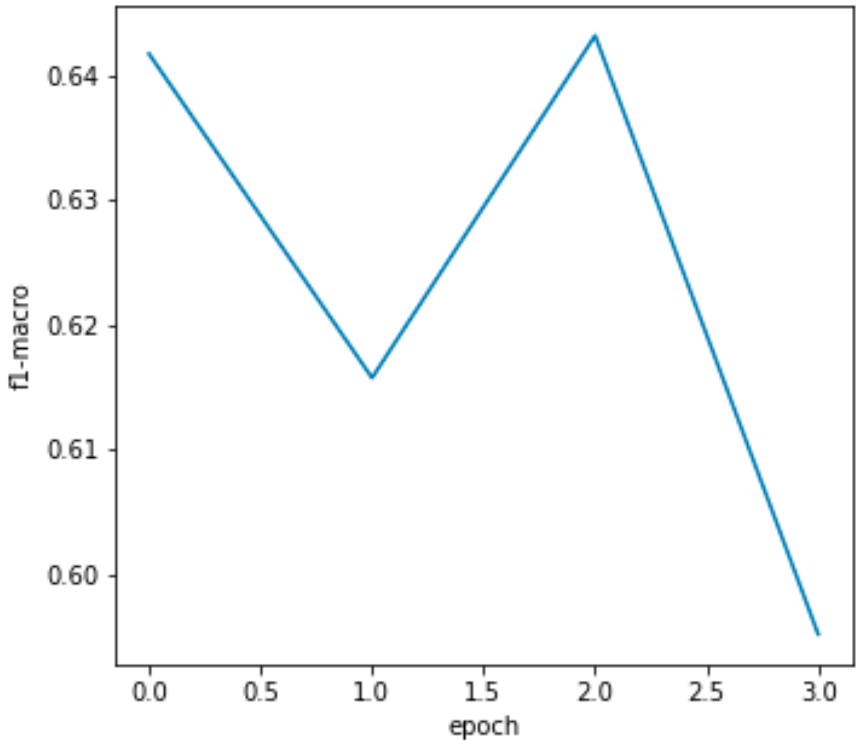
معیارهای ارزیابی بر روی داده‌های ارزیابی به شرح زیر است.

metrics	Test
F1-score macro	0.624
F1-score micro	0.803
Accuracy score	0.803
Recall score	0.383
ROCAUC score	0.629

مدل Roberta

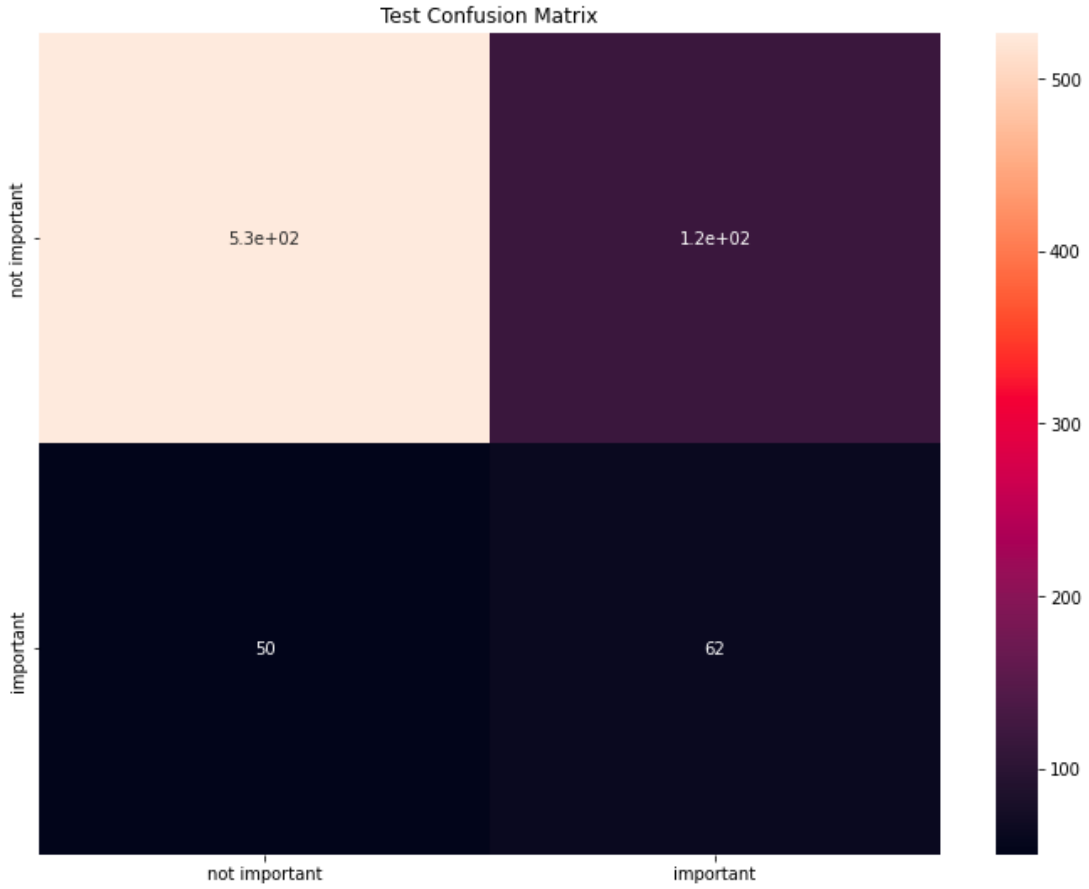
نمودار هزینه و معبار ارزیابی f1-macro در زمان آموزش





مدل Roberta

معیارهای ارزیابی بر روی داده‌های ارزیابی به شرح زیر است.



metrics	Test
F1-score macro	0.643
F1-score micro	0.778
Accuracy score	0.778
Recall score	0.553
Precision score	0.344
ROCAUC score	0.685

استخراج کلمات کلیدی مهم برای دسته اخبار مهم

درفایل Roberta_News_Final.ipynb این گام انجام شده است و برای 40 عدد از کلمات کلیدی در هر دسته نتایج قابل مشاهده هستند که بعضی از آنها به صورت زیر می‌باشند.

```
(0.028756668642294647 ,'هخامنشی ')
(b', 0.029025569642145532')
(chart', 0.030962146370082433')
(0.030069965087503148 ,'اختیارتان ')
(0.02931130029732487 ,'بته')
(0.02913316623218165 ,'کوجیما ')
(0.02880508646949198 ,'مارش ')
(0.029003703361570654 ,'ترویج ')
(0.030074425659835757 ,'حساب')
(0.0299403362727384 ,'لطف ')
(0.0304519749398533 ,'جوانب ')
(0.02942515381675448 ,'اریوم')
(0.031019451281345756 ,'رندوم ')
(Lenovo', 0.03135095601779092 ')
(0.03245496617018051 ,'ابه ')
(0.034282009333890695 ,'ماندلا ')
(0.03218991110407354 ,'مزمن ')
(0.0334688739796698 ,'دستکاری ')
```

```
(0.026857626524902756 ,'تمایلی ')
(0.026946033823686433 ,'ریلکس ')
(0.02710694842803396 ,'لابی ')
(0.027850745707358904 ,'زبانه')
(0.02830167646036852 ,'صرافی ')
(0.02748162277821975 ,'کریس')
(0.027262430045943198 ,'فکر')
(0.02811200028148786 ,'رهنورد ')
(0.02865137406680507 ,'یهود ')
(0.02715795591264758 ,'کسر ')
(0.028168105824208922 ,'بپرهیزید ')
(0.027499112149240945 ,'مناظری ')
(0.02791745213786756 ,'وپاتی')
(0.028564986371280243 ,'لارنس ')
(0.027396756352601415 ,'فحش ')
(0.028390117493986233 ,'ویت ')
(0.02769914886078828 ,'جوید ')
(0.028546214779351597 ,'حلیم ')
```

استخراج کلمات کلیدی مهم برای دسته اخبار غیرمهم

درفایل Roberta_News_Final.ipynb این گام انجام شده است و برای 40 عدد از کلمات کلیدی در هر دسته نتایج قابل مشاهده هستند که بعضی از آنها به صورت زیر می‌باشند.

```
(0.026250738243382088 ,' حنفی ')
(0.02706729912392758 ,' باقیم ')
(0.02673050648383901 ,' ازلحاظ ')
(0.027018043632219158 ,' جادوگری '
(uel', 0.027606222265781355')
(0.027066850279600985 ,' شکسته ')
(0.028444070068159633 ,'استیون')
(0.027049394225613144 ,' باربر ')
(0.02648436982560165 ,' تراک ')
(0.02741041703054381 ,' استیصال ')
(ight', 0.02814076747479266')
(0.027072297749678675 ,' نیجریه ')
(0.027209425261396802 ,' وسیله ')
(0.0275377827823704 ,' برو ')
(0.027026747715776622 ,' سیروس ')
(0.027320303422406544 ,'پوشه')
(0.02731604513639585 ,' لوبیای ')
(0.027403846953560462 ,' بستری ')
(0.02825394744851817 ,' درود ')
```

```
(0.02650189137496861 ,' چگ ')
(0.028606389073085436 ,' ماراتن ')
(place', 0.0309548802930621 48')
(AND', 0.0289318917088876 58')
(0.030813286853133226 ,' میزان ')
(0.0310639277040623 ,' روایی ')
(0.029059050414706133 ,' زرد ')
(0.029410773213123464 ,'شخصی')
(0.0314211251295895 ,' سریانی ')
(0.032227744249300194 ,'پاری')
(anced', 0.0305415490539233 67')
(CMOS', 0.028965734146189 97 ')
(0.028638508446543032 ,' بلوتوثی '
(0.029138713947217676 ,'فاصله')
(0.028882930419912523 ,' قرنیه ')
(0.03144761721795253 ,' مخمر ')
(0.029799696487942562 ,' توجه ')
(0.02932936335420755 ,' معنای ')
```

جمع بندی تاثیر بخش‌های مختلف بر مدل‌های کلاسیک

با توجه به معیار f1-macro بعنوان جمع بندی برای این قسمت خواهیم داشت.

	Logistic regression	svm
Text	0.649	0.649
Text + title	0.630	0.629
Text + category	0.621	0.627
Text + keywords	0.627	0.628
Text + remove source name	0.646	0.647

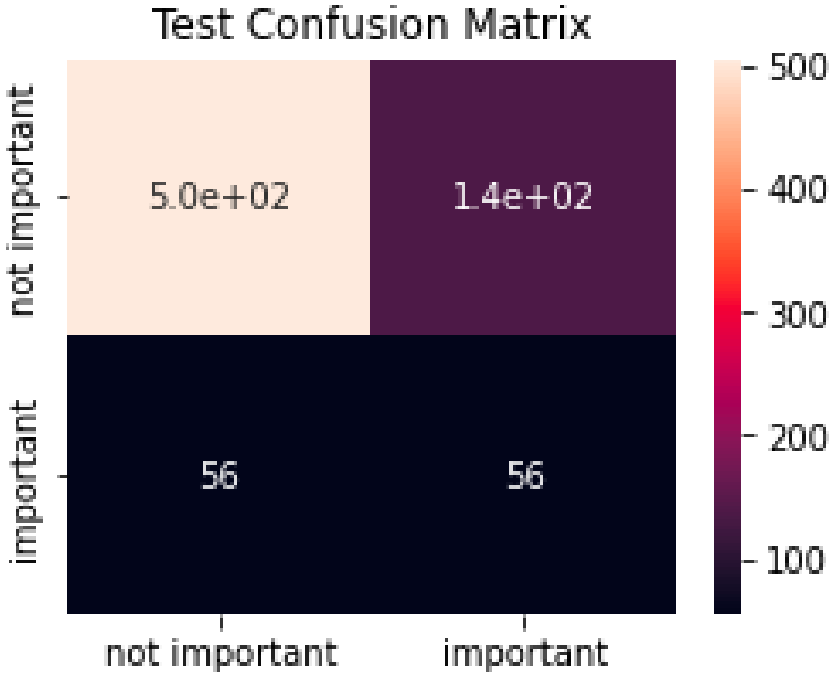
تاثیر اعمال کلمات کلیدی و دسته خبر بر مدل Roberta

metrics	Text	Text + Category	Text + Keywords
F1-score macro	0.643	0.668	0.626
F1-score micro	0.778	0.812	0.813
Accuracy score	0.778	0.812	0.813
Recall score	0.553	0.517	0.357
Precision score	0.344	0.3	0.3
ROCAUC score	0.685	0.690	0.625

استفاده از k-means جهت انتخاب اخبار غیر تاثیرگذار

از مدل Roberta برای بررسی این مورد استفاده شده است و نتایج آموزش به شرح زیر می‌باشند.

معیارهای ارزیابی بر روی داده‌های تست به شرح زیر است.

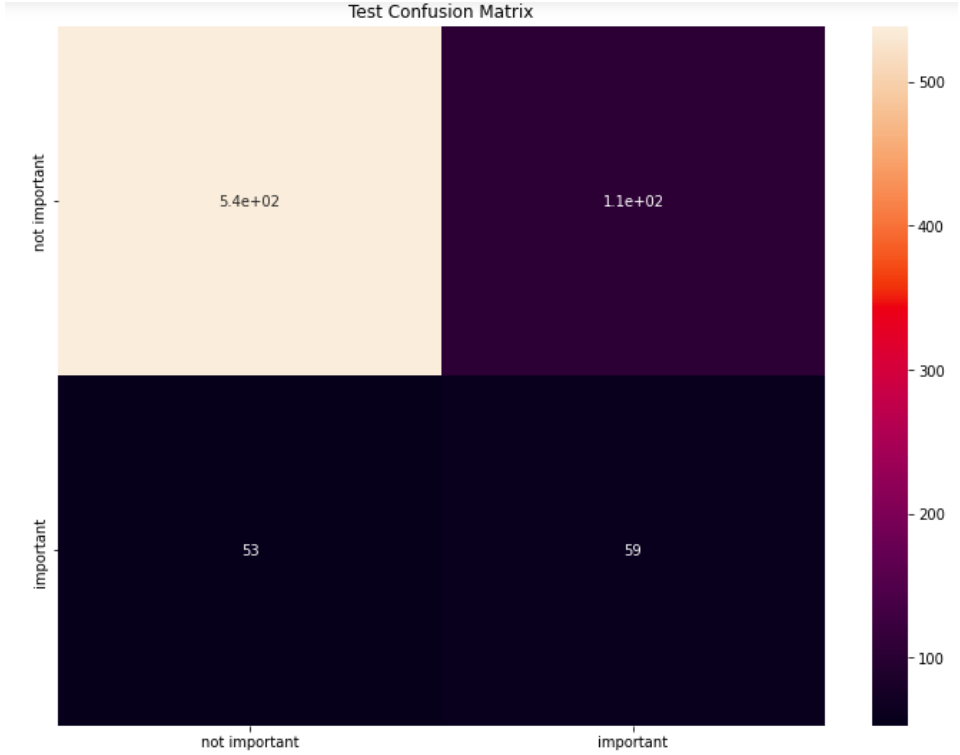


metrics	original	k-means
F1-score macro	0.643	0.600
F1-score micro	0.778	0.741
Accuracy score	0.778	0.741
Recall score	0.553	0.5
Precision score	0.344	0.28
ROCAUC score	0.685	0.641

استفاده از Oversampling برای Roberta

از مدل Roberta برای بررسی این مورد استفاده شده است و نتایج آموزش به شرح زیر می‌باشند.

معیارهای ارزیابی بر روی داده‌های ارزیابی به شرح زیر است.



metrics	original	Oversampling
F1-score macro	0.643	0.648
F1-score micro	0.778	0.789
Accuracy score	0.778	0.789
Recall score	0.553	0.526
Precision score	0.344	-
ROCAUC score	0.685	0.681

تاثیر بایاس بودن نسبت به نام خبر گذاری

با توجه به معیار f1-macro بعنوان جمع بندی برای این قسمت خواهیم داشت.

	Text	Text – Source Name
Logistic Regression	64.9	64.6
SVM	64.9	64.7
Roberta	0.643	0.639
LSTM-attention	0.623	0.624

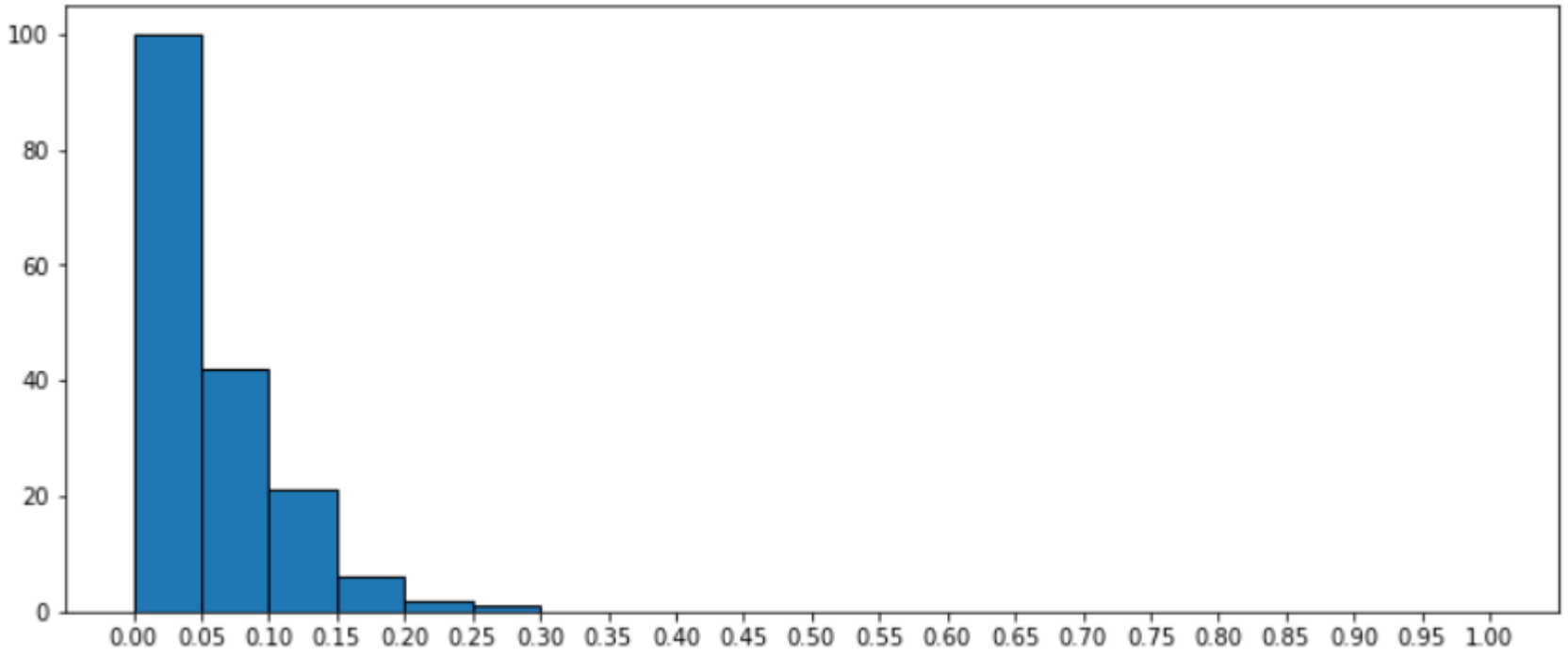
نتایج correlation دسته خبر با اخبار مهم

❖ نتایج کامل مدل‌های دیگر یعنی NN و Roberta در فایل scorings.py قابل مشاهده است. (محاسبه شده توسط mutual information)

```
Logistic Regression Category Mutual Information
0.029697166109571738 <-- اجتماعی
0.026101939468282875 <-- اقتصادی
0.023554850151957796 <-- بین الملل
0.018625982350700676 <-- حوادث
0.010467847370794425 <-- سیاسی
0.01435428777387604 <-- علمی و پزشکی
0.03302084623623136 <-- فرهنگ و هنر
0.01595202012622865 <-- فناوری و ارتباطات
0.009801944831982379 <-- مذهبی
0.0222053727520239 <-- ورزشی
```

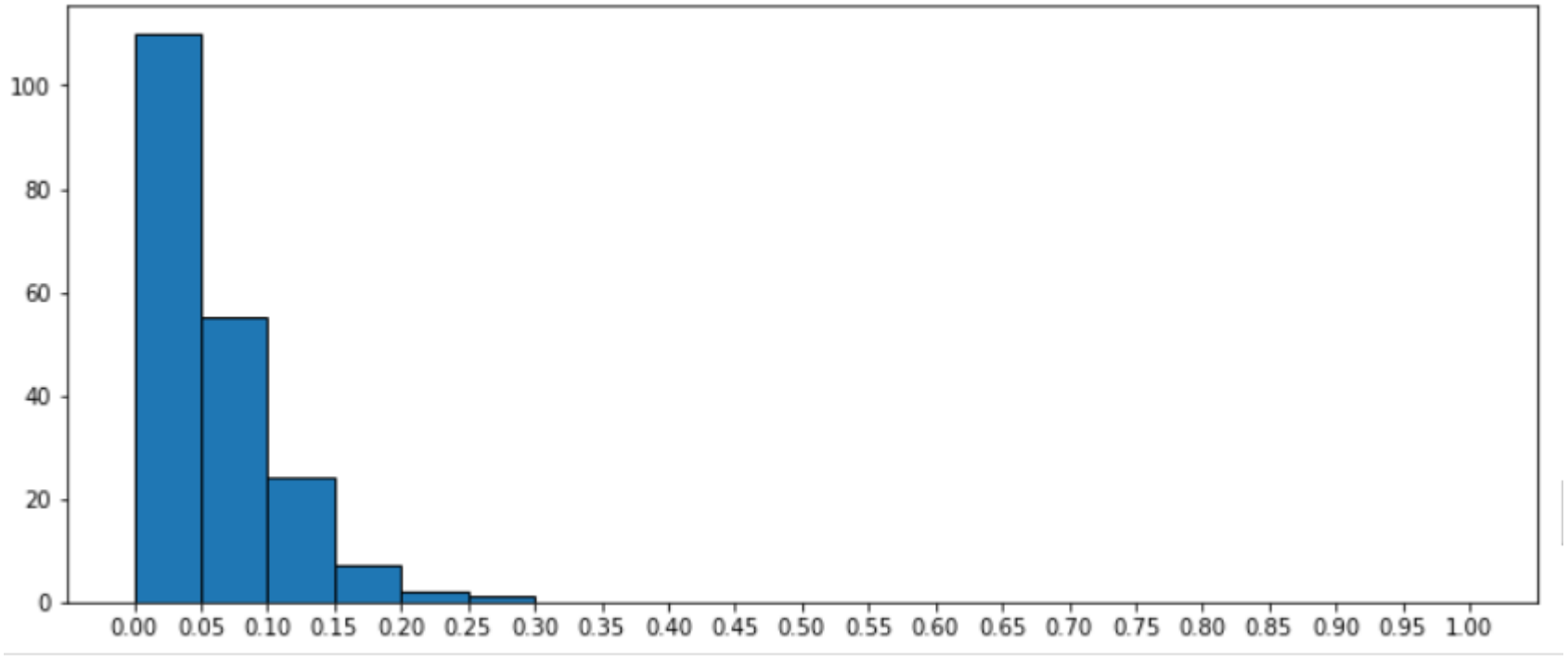
نتایج توزیع خطای دیتای اشتباه

❖ Logistic Regression



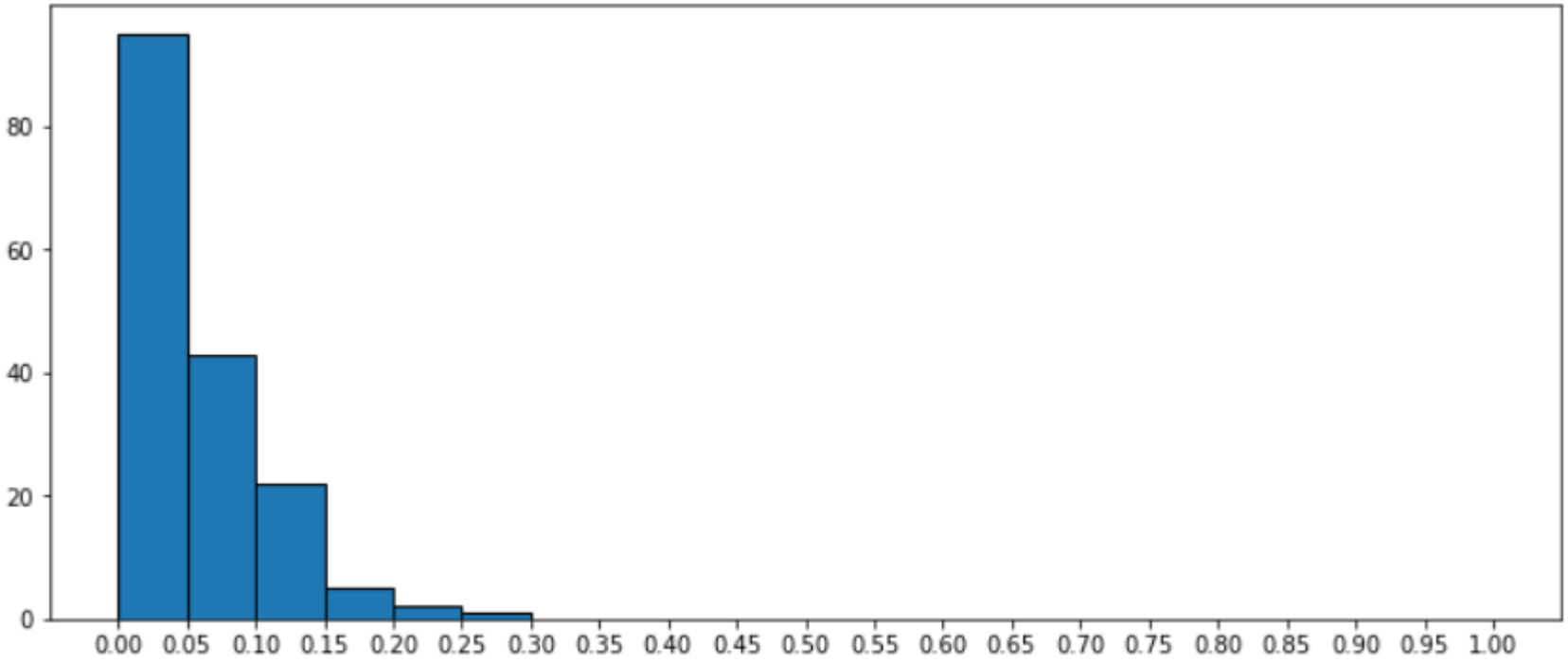
نتایج توزیع خطای دیتای اشتباه

Neural Network ❖



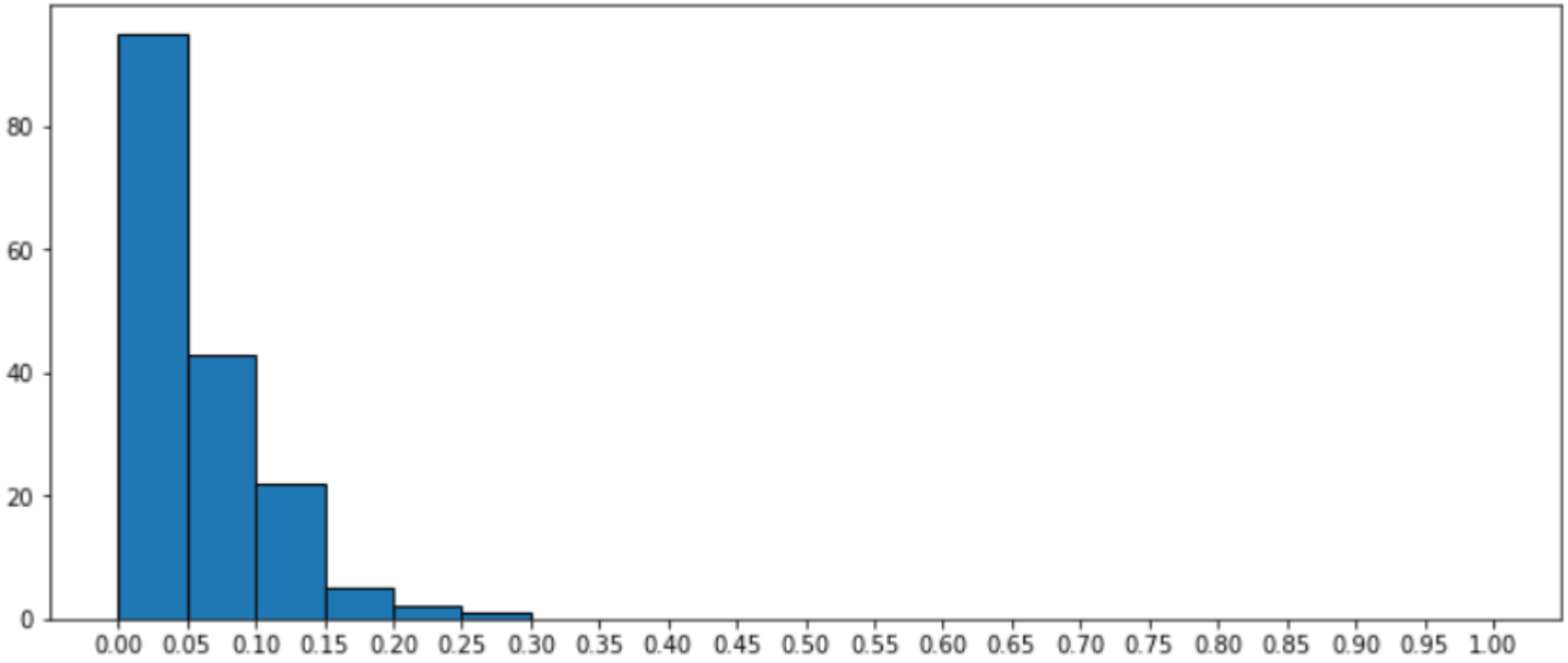
نتایج توزیع خطای دیتای اشتباه

- Roberta



نتایج توزیع خطای دیتای اشتباه

- Roberta



مقایسه نهایی f1-macro بر روی چند مدل

Model	F1-Macro
LR	0.649
SVM	0.649
Bilstm	0.624
Roberta + category	**0.668**
Roberta + KMeans	0.600
Roberta + Oversampling	0.648
LR + Bilstm + Roberta	**0.675***

نتیجه گیری و جمع بندی

- بهترین نتیجه مدل ensemble با f1-macro برابر 67.5 بود.
- بسیاری از داده‌هایی که اشتباه دسته‌بندی میشوند درصد خطای پایینی دارند.
- مدلی که روی داده ما آموزش دیده روی مدل ۲ گروه دیگر نتایج تقریبا مشابهی میداد.
- پس از آزمایش روی تمامی داده‌های گروه‌ها دقت بهبود نیافت و دقت حوالی دقت با استفاده از دیتای خودمان بود.

پیوست

- در این قسمت نتایج بیشتر پروژه قابل مشاهده است.

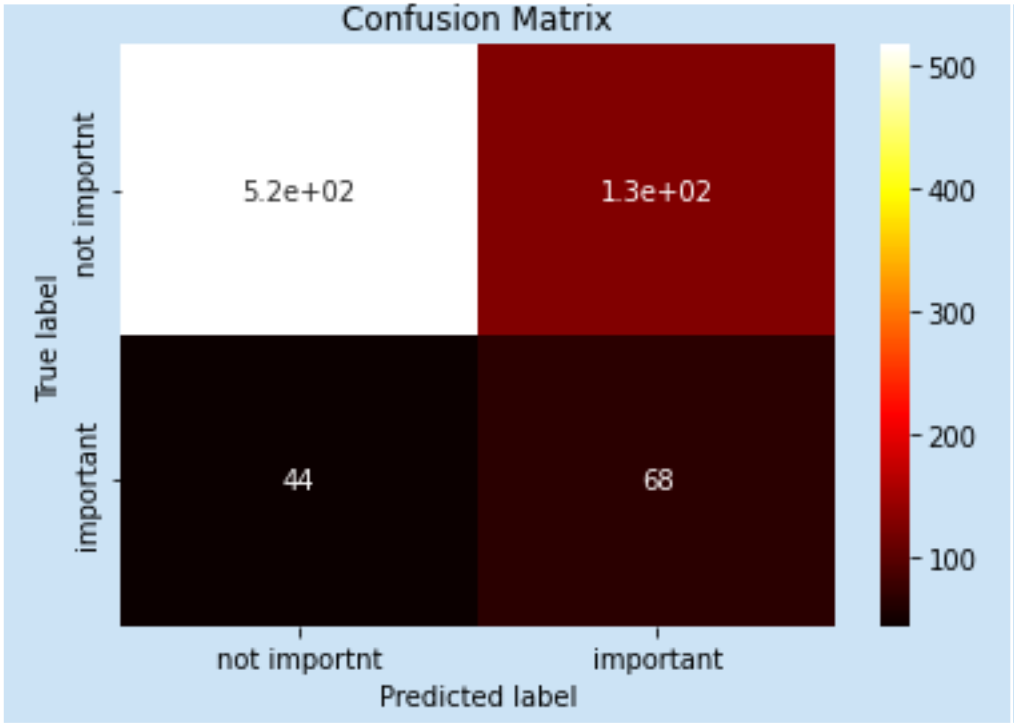
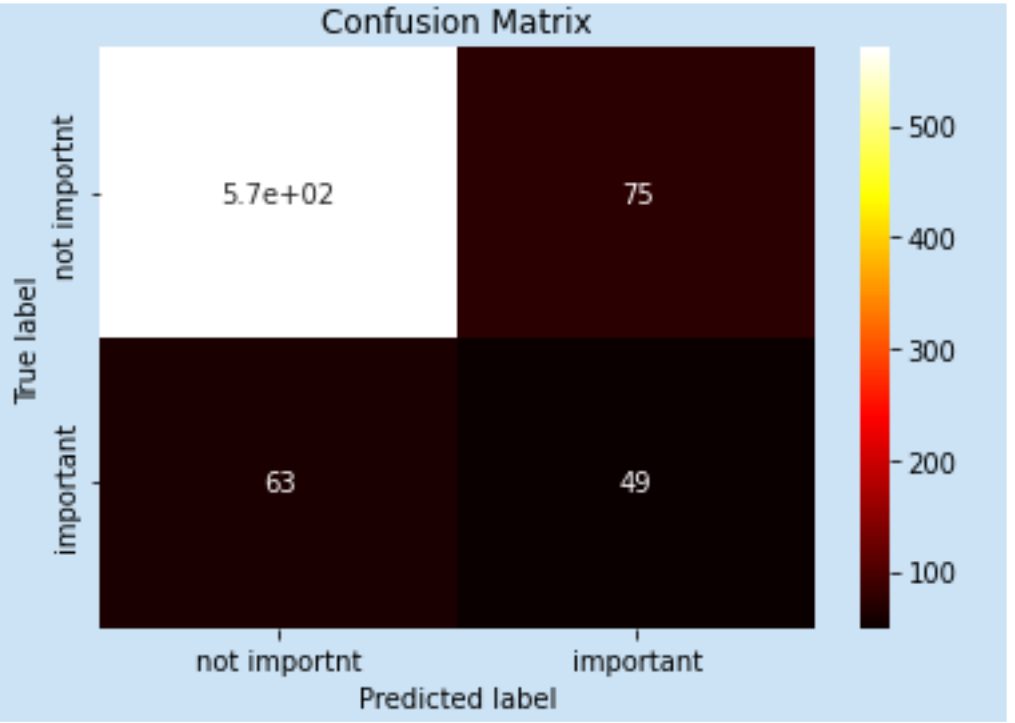
مدل Naive Bayes(Multinomial)

- ✓ از cross-validation با c=5 جهت آموزش مدل و یافتن بهترین پارامترها استفاده شده است.
- ✓ زمان آموزش حدودا 5 دقیقه.

metrics	Before tuning		After tuning	
	TF-IDF	BoW	TF-IDF	BoW
F1-score macro	0.460	0.503	0.460	0.503
F1-score micro	0.852	0.857	0.852	0.857
Accuracy score	0.852	0.857	0.852	0.857
Recall score	0.5	0.521	0.5	0.521
Precision score	0.426	0.845	0.426	0.845
ROCAUC score	0.5	0.512	0.5	0.512
Grid Search parameters	-	-	Alpha=1	Alpha=1

مدل Naive Bayes(Multinomial)

- ✓ ماتریس درهم ریختگی برای مدل fine tune شده.

TF-IDF	BoW
	

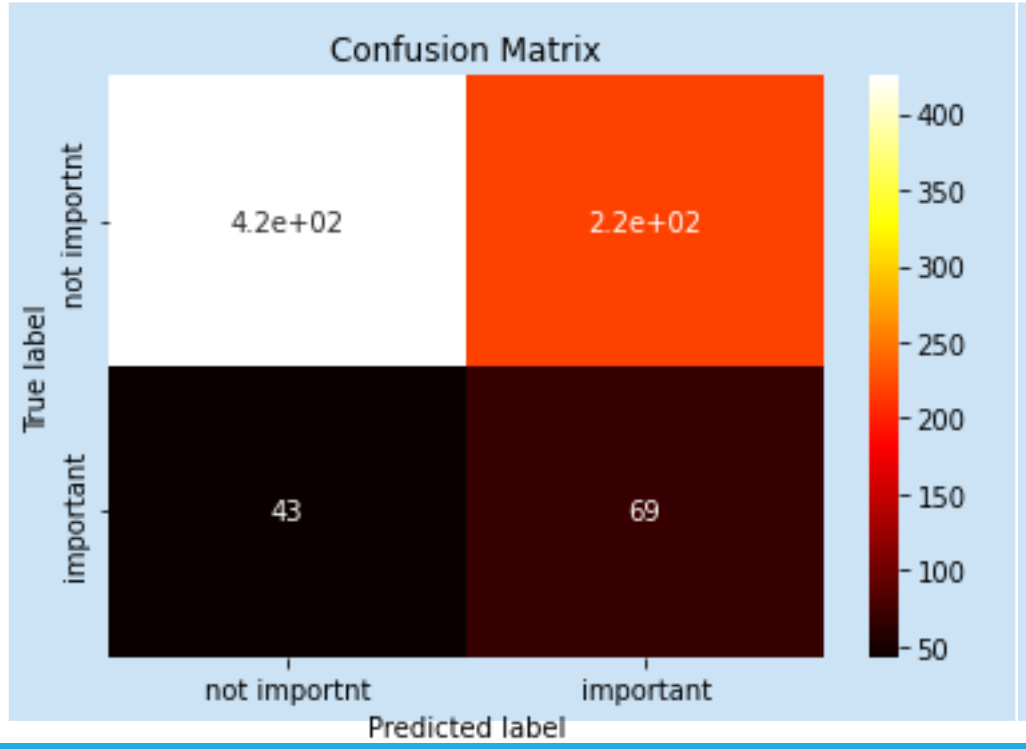
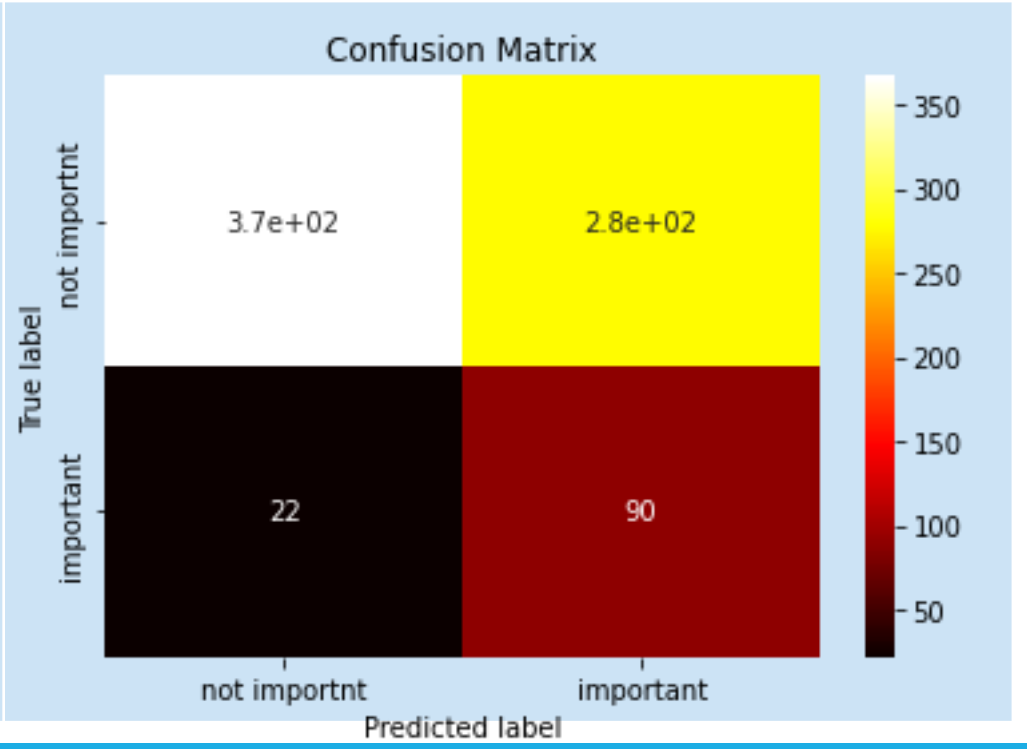
مدل Random Forest

- ✓ از cross-validation با c=5 جهت آموزش مدل و یافتن بهترین پارامترها استفاده شده است.
- ✓ زمان آموزش حدودا ۵ دقیقه.

metrics	Before tuning		After tuning	
	TF-IDF	BoW	TF-IDF	BoW
F1-score macro	0.459	0.468	0.553	0.542
F1-score micro	0.850	0.852	0.652	0.603
Accuracy score	0.850	0.852	0.652	0.603
Recall score	0.499	0.503	0.637	0.686
Precision score	0.425	0.676	0.573	0.594
ROCAUC score	0.499	0.503	0.637	0.686
Grid Search parameters	-	-	class weight= {0: 0.15, 1: 1}, criterion= gini, max_depth= 3	

مدل Random Forest

- ✓ ماتریس درهم ریختگی برای مدل fine tune شده.

TF-IDF	BoW
	

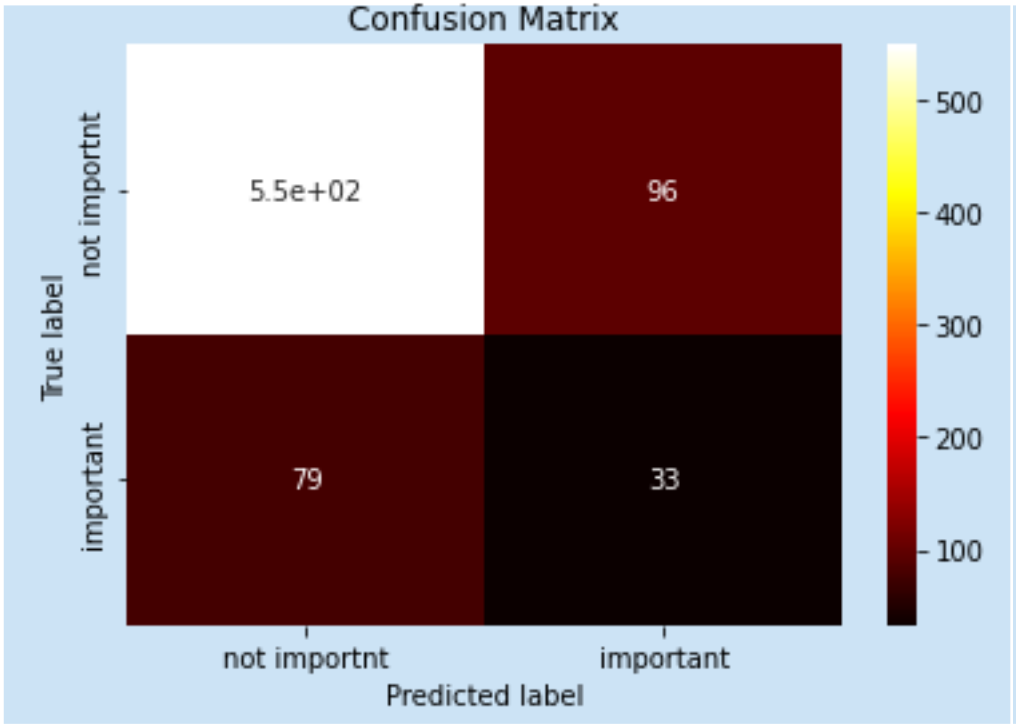
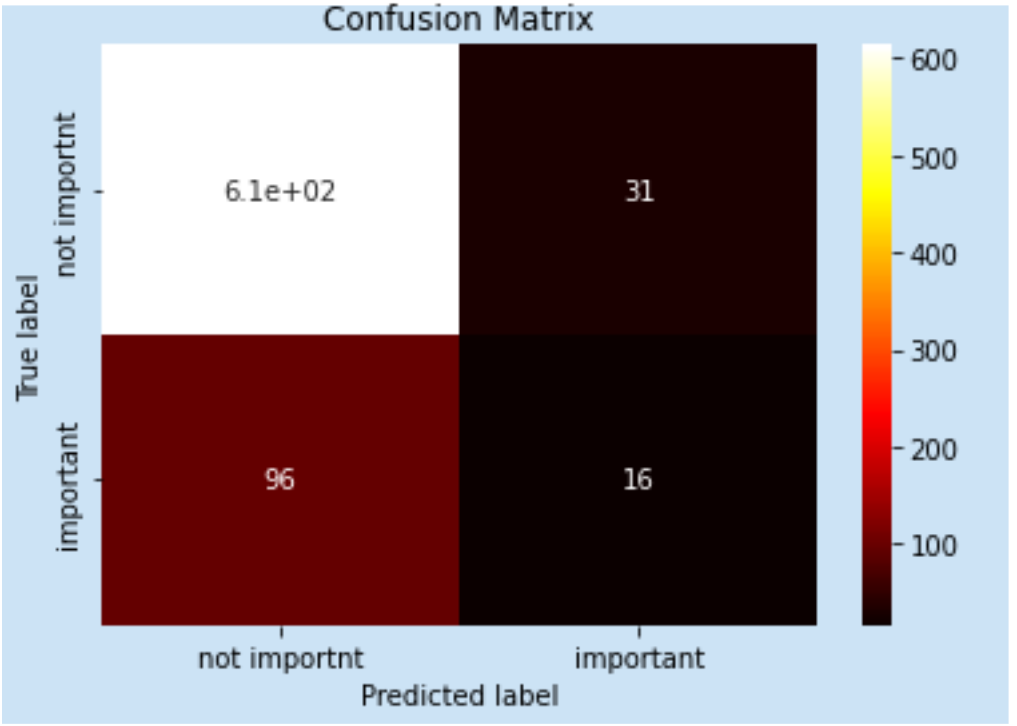
مدل Decision Tree

- ✓ از cross-validation با c=5 جهت آموزش مدل و یافتن بهترین پارامترها استفاده شده است.
- ✓ زمان آموزش حدودا ۵ دقیقه.

metrics	Before tuning		After tuning	
	TF-IDF	BoW	TF-IDF	BoW
F1-score macro	0.559	0.543	0.568	0.553
F1-score micro	0.795	0.829	0.768	0.832
Accuracy score	0.795	0.829	0.768	0.832
Recall score	0.577	0.571	0.572	0.547
Precision score	0.582	0.618	0.565	0.602
ROCAUC score	0.577	0.571	0.572	0.547
Grid Search parameters	-	-	Criterion=gini , splitter=random, Max depth= 3, Class weight={0:0.3, 1:1}	

مدل Decision Tree

- ✓ ماتریس درهم ریختگی برای مدل fine tune شده.

TF-IDF	BoW
	

بررسی تاثیر سایر ویژگی‌ها بر کارایی مدل‌های کلاسیک

جهت انجام این بررسی از ویژگی‌های دیگری از جمله عنوان خبر، کلمات کلیدی و دسته خبر نیز استفاده می‌کنیم.

توزیع داده‌های ارزیابی در مجموعه دادگان ارزیابی به شکل زیر است:

دسته	تعداد کل	تعداد خبر مهم	درصد
بین الملل	177	27	23.6
اقتصادی	81	20	17.5
حوادث	51	8	7
اجتماعی	68	9	7.8
مذهبی	58	1	0.8
علمی و پزشکی	68	19	16.6
ورزشی	118	11	9.6
فرهنگ و هنر	68	2	1.7
فناوری و ارتباطات	54	11	9.6
سیاسی	74	6	5.2

تاثیر عنوان (متن + عنوان)

برای بررسی اهمیت در نظر گرفتن عنوان با استفاده از دو مدل تیون شده این بررسی را انجام می‌دهیم.

metrics	Logistic Regression	SVM
	TF-IDF	TF-IDF
F1-score macro	0.630	0.629
F1-score micro	0.770	0.762
Accuracy score	0.770	0.762
Recall score	0.666	0.672
Precision score	0.618	0.617
ROCAUC score	0.666	0.672

تاثیر دسته خبر(متن + دسته خبر)

برای بررسی اهمیت در نظر گرفتن دسته خبر با استفاده از دو مدل تیون شده این بررسی را انجام می‌دهیم.

metrics	Logistic Regression	SVM
	TF-IDF	TF-IDF
F1-score macro	0.621	0.627
F1-score micro	0.763	0.763
Accuracy score	0.763	0.763
Recall score	0.655	0.655
Precision score	0.609	0.614
ROCAUC score	0.655	0.655

تاثیر کلمات کلیدی (متن + کلمات کلیدی)

برای بررسی اهمیت در نظر گرفتن کلمات کلیدی با استفاده از دو مدل تیون شده این بررسی را انجام می‌دهیم.

metrics	Logistic Regression	SVM
	TF-IDF	TF-IDF
F1-score macro	0.627	0.628
F1-score micro	0.768	0.763
Accuracy score	0.768	0.763
Recall score	0.661	0.669
Precision score	0.615	0.616
ROCAUC score	0.661	0.669

تاثیر عنوان (متن + عنوان)

توزیع خروجی‌های مدل logistic regression.

دسته	تعداد کل	تعداد خبر مهم	درصد
بین الملل	177	38	21.3
اقتصادی	81	38	21.3
حوادث	51	0	0
اجتماعی	68	11	6.1
مذهبی	58	1	0.6
علمی و پزشکی	68	35	19.6
ورزشی	118	27	15.16
فرهنگ و هنر	68	0	0
فناوری و ارتباطات	54	11	6.1
سیاسی	74	17	9.5

تاثیر دسته خبر (متن + دسته خبر)

توزیع خروجی‌های مدل SVM.

دسته	تعداد کل	تعداد خبر مهم	درصد
بین الملل	177	37	20
اقتصادی	81	36	19.45
حوادث	51	2	1
اجتماعی	68	10	5.4
مذهبی	58	1	0.5
علمی و پزشکی	68	37	20
ورزشی	118	28	15.15
فرهنگ و هنر	68	0	0
فناوری و ارتباطات	54	19	10.2
سیاسی	74	15	8.1

تاثیر کلمات کلیدی (متن + کلمات کلیدی)

توزیع خروجی‌های مدل SVM.

دسته	تعداد کل	تعداد خبر مهم	درصد
بین الملل	177	37	19.7
اقتصادی	81	35	18.7
حوادث	51	2	1
اجتماعی	68	11	5.8
مذهبی	58	2	1
علمی و پزشکی	68	35	18
ورزشی	118	29	15.5
فرهنگ و هنر	68	0	0
فناوری و ارتباطات	54	19	10.1
سیاسی	74	17	9

جمع بندی تاثیر بخش‌های مختلف برمدل‌های کلاسیک (درصد تاثیرگذاری)

دسته	برچسب‌های اصلی	متن + عنوان	متن + دسته خبر	متن + کلمات کلیدی
بین الملل	23.6	21.3	20	19.7
اقتصادی	17.5	21.3	19.45	18.7
حوادث	7	0	1	1
اجتماعی	7.8	6.1	5.4	5.8
مذهبی	0.8	0.6	0.5	1
علمی و پزشکی	16.6	19.6	20	18
ورزشی	9.6	15.16	15.15	15.5
فرهنگ و هنر	1.7	0	0	0
فناوری و ارتباطات	9.6	6.1	10.2	10.1
سیاسی	5.2	9.5	8.1	9