Midterm Report
Lingwei, Xiaorui
Feburary 18th, 2016

Progress Track

"By the time of midterm presentation, we will …"

- Finish literature review: We did some before the proposal but we realized that we did not do it adequately and need more information to refine our approach.
- Data collection: Rather than 200 articles from each category we downloaded 100 from each.
- Preprocessing: We did some basic web scraping and we need to clean the information more.
- Feature generation: We generated the features we listed in our previous report including POS which include modifiers and tenses, TF-IDF, Negation, Bi-gram.
- First pass of the model and output the evaluation results: See below.
- Develop a basic version of user interface: Used flask for a proof of concept

Adaptation and Obstacles

- Although the evaluation results are pretty good, they probably reflect the differences between the writing styles of USAToday and FactCheck, rather than the real difference in facts and opinions.
- The bags of words generated from the html objects are not clean enough, containing JavaScript codes. The get_words() function needs to be revised and tested.
- The models can generate seemly rather accurate results for long texts, but they cannot classify short texts reliably. To explore this, we can generate plot of text length vs correct classification.
- Feature generation: We originally intended to generate separate features for modifiers and past tenses as they were suggested as good indicators for distinguishing opinions from news. However, this task seems trivial given part-of-speech tags which already distinguish verbs into present/past tense and also have three forms for adjectives and adverbs. We think generating separate features for them will duplicate the POS results. For now, we are looking for advice on what other useful features could we produce.
- User interface: Without a powerful server, it took significant time to train the classifier.

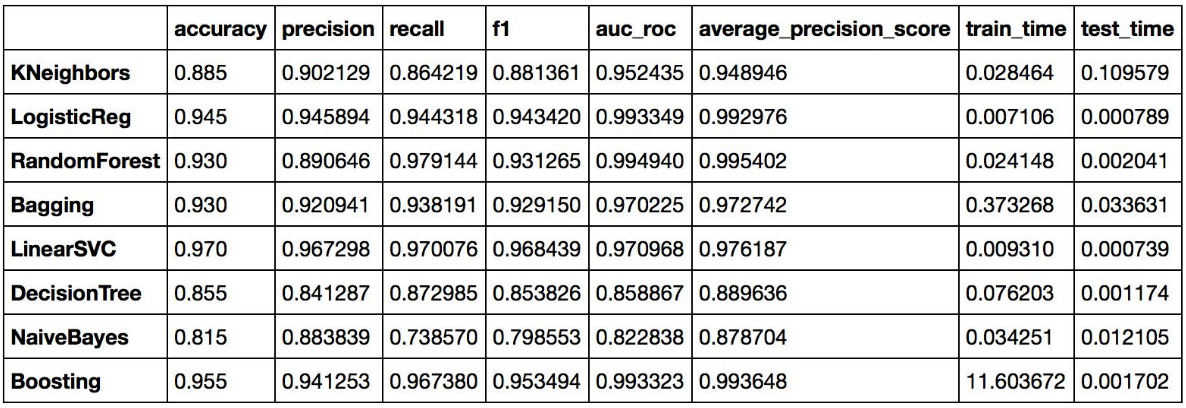
Next Steps

- Improve the preprocessing functions to generate better web scrape output: ignore quotes and javascript code.
- Add more features to the models: this requires us to discuss with the professor and conduct more literature review. We need to find a good approach to generate features with opinion seed words. We are considering using stemmer and wordnet for more features.

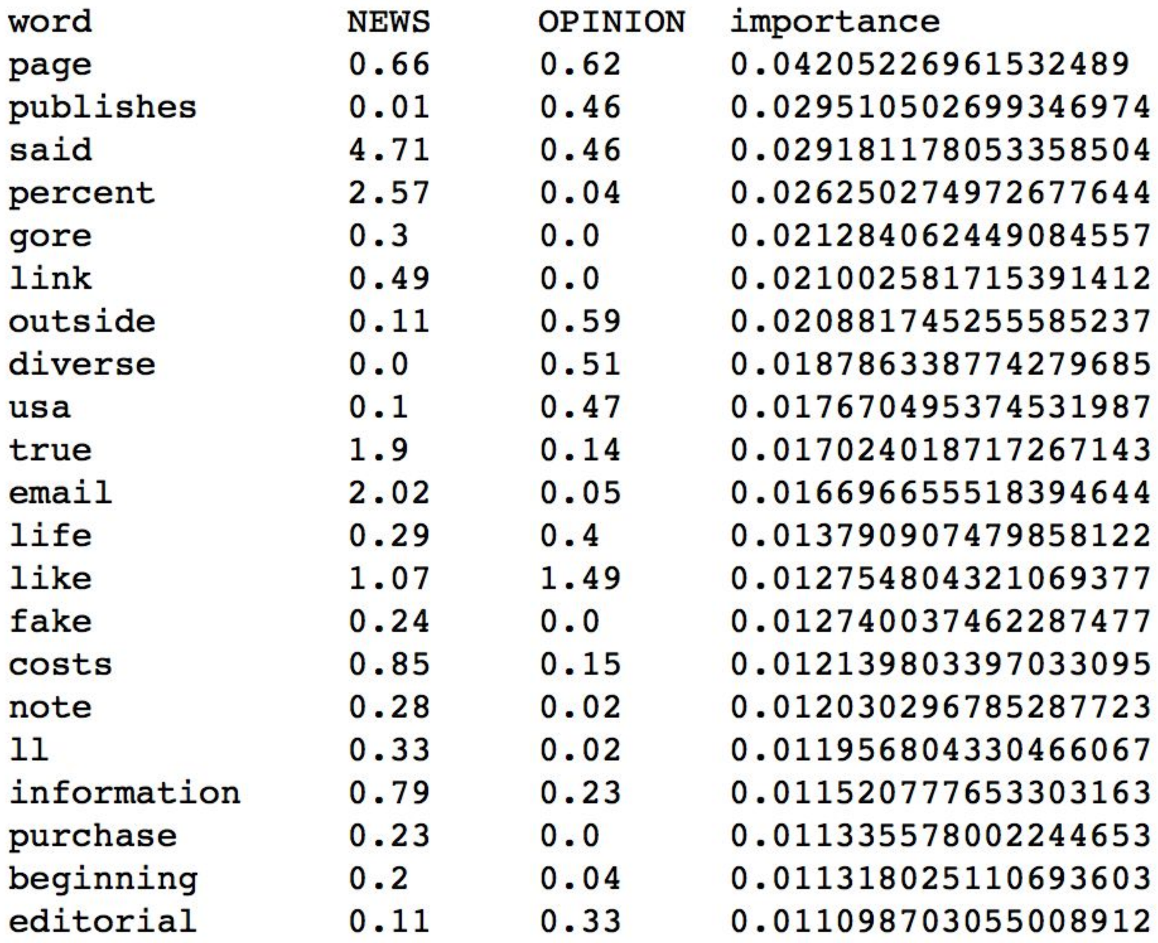
- Implement smoothing techniques: we are not clear about how to implement this yet.
- Implement feature selection: we currently don't have that many features that require us to select a few but this might become a need as we generate more features.
- Improve the design and functionality of the user interface

Evaluation Results & Important Features

1) Unigrams

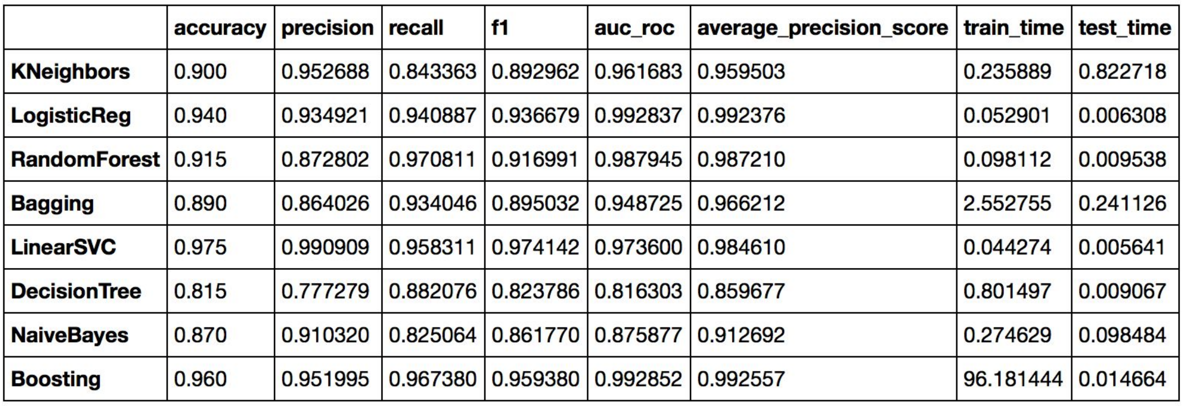


	accuracy	precision	recall	f1	auc_roc	average_precision_score	train_time	test_time
KNeighbors	0.885	0.902129	0.864219	0.881361	0.952435	0.948946	0.028464	0.109579
LogisticReg	0.945	0.945894	0.944318	0.943420	0.993349	0.992976	0.007106	0.000789
RandomForest	0.930	0.890646	0.979144	0.931265	0.994940	0.995402	0.024148	0.002041
Bagging	0.930	0.920941	0.938191	0.929150	0.970225	0.972742	0.373268	0.033631
LinearSVC	0.970	0.967298	0.970076	0.968439	0.970968	0.976187	0.009310	0.000739
DecisionTree	0.855	0.841287	0.872985	0.853826	0.858867	0.889636	0.076203	0.001174
NaiveBayes	0.815	0.883839	0.738570	0.798553	0.822838	0.878704	0.034251	0.012105
Boosting	0.955	0.941253	0.967380	0.953494	0.993323	0.993648	11.603672	0.001702

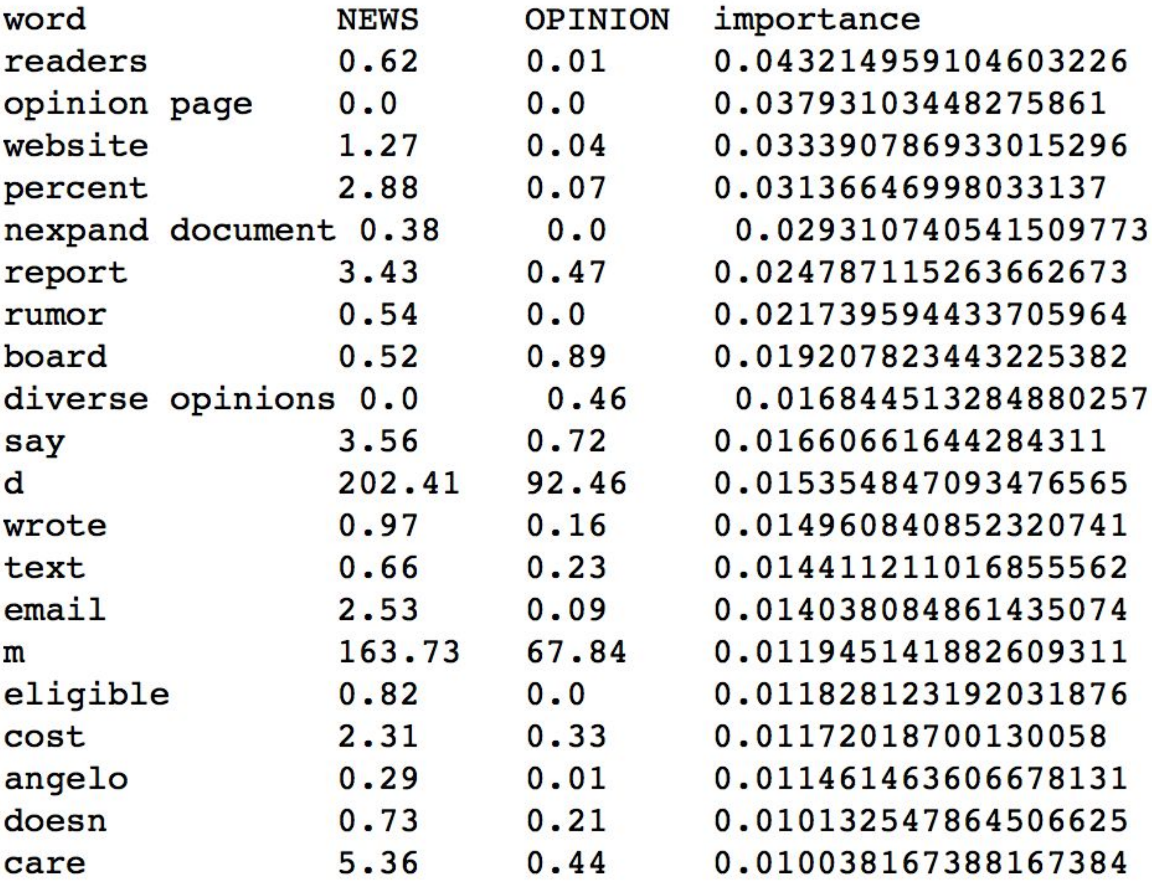


```
word                NEWS       OPINION    importance
page                0.66       0.62       0.04205226961532489
publishes           0.01       0.46       0.029510502699346974
said                4.71       0.46       0.029181178053358504
percent             2.57       0.04       0.026250274972677644
gore                0.3        0.0        0.021284062449084557
link                0.49       0.0        0.021002581715391412
outside             0.11       0.59       0.020881745255585237
diverse             0.0        0.51       0.018786338774279685
usa                 0.1        0.47       0.017670495374531987
true                1.9        0.14       0.017024018717267143
email               2.02       0.05       0.016696655518394644
life                0.29       0.4        0.013790907479858122
like                1.07       1.49       0.012754804321069377
fake                0.24       0.0        0.012740037462287477
costs               0.85       0.15       0.012139803397033095
note                0.28       0.02       0.012030296785287723
ll                  0.33       0.02       0.011956804330466067
information         0.79       0.23       0.011520777653303163
purchase            0.23       0.0        0.011335578002244653
beginning           0.2        0.04       0.011318025110693603
editorial           0.11       0.33       0.011098703055008912
```

2) Unigrams and Bigrams

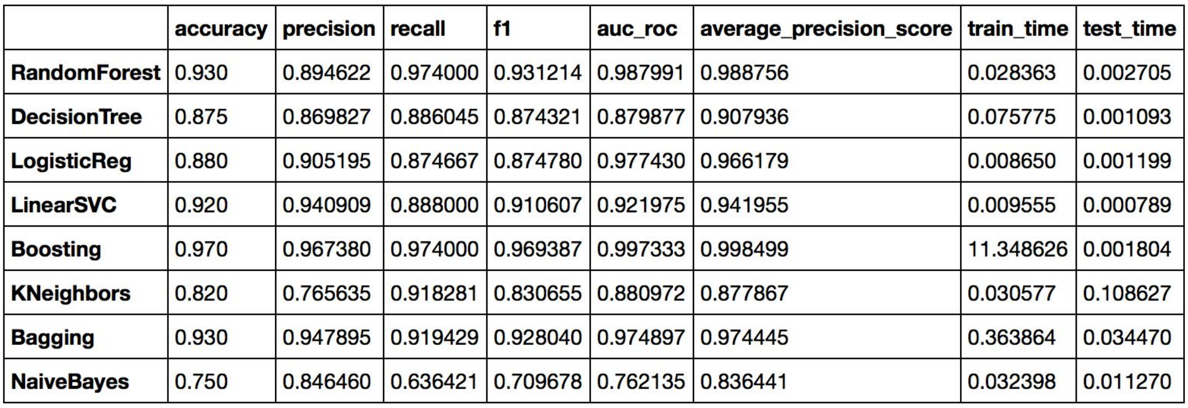


	accuracy	precision	recall	f1	auc_roc	average_precision_score	train_time	test_time
KNeighbors	0.900	0.952688	0.843363	0.892962	0.961683	0.959503	0.235889	0.822718
LogisticReg	0.940	0.934921	0.940887	0.936679	0.992837	0.992376	0.052901	0.006308
RandomForest	0.915	0.872802	0.970811	0.916991	0.987945	0.987210	0.098112	0.009538
Bagging	0.890	0.864026	0.934046	0.895032	0.948725	0.966212	2.552755	0.241126
LinearSVC	0.975	0.990909	0.958311	0.974142	0.973600	0.984610	0.044274	0.005641
DecisionTree	0.815	0.777279	0.882076	0.823786	0.816303	0.859677	0.801497	0.009067
NaiveBayes	0.870	0.910320	0.825064	0.861770	0.875877	0.912692	0.274629	0.098484
Boosting	0.960	0.951995	0.967380	0.959380	0.992852	0.992557	96.181444	0.014664

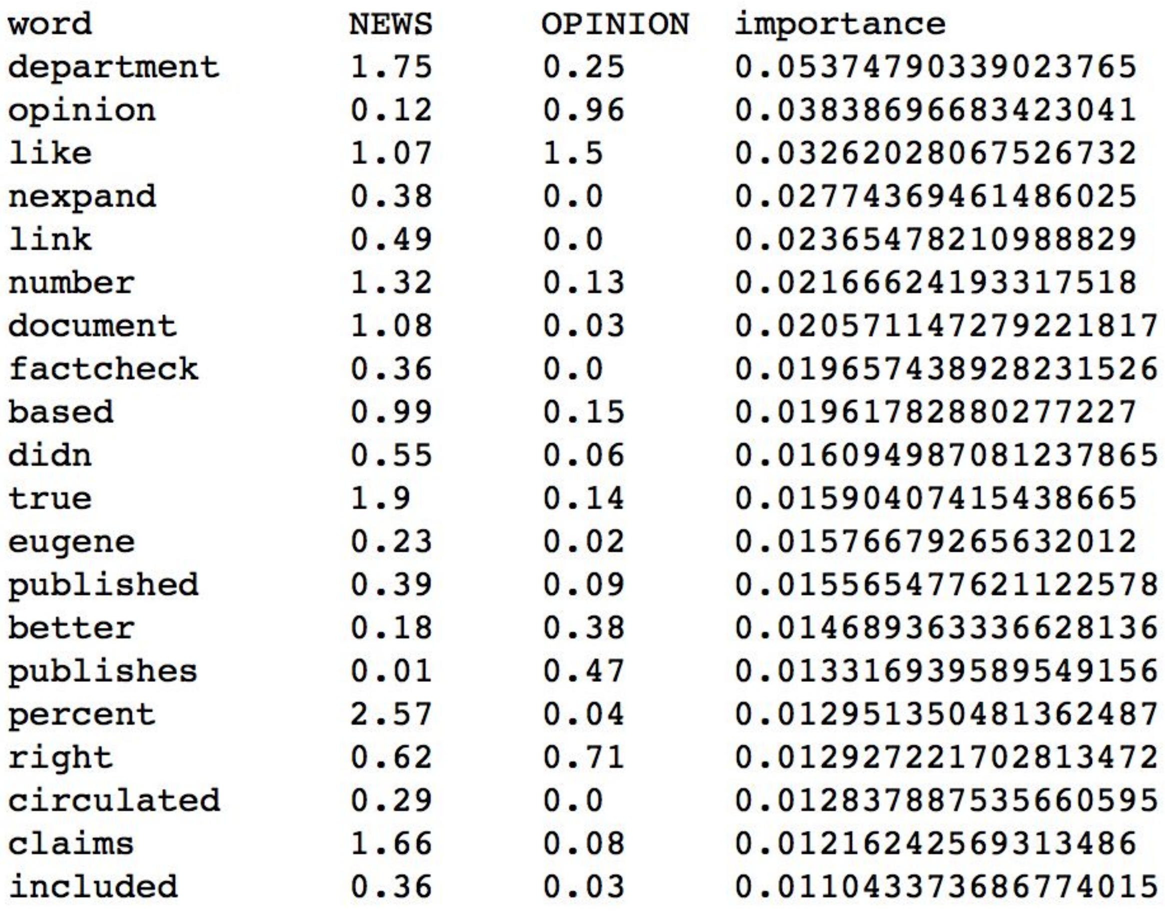


```
word            NEWS     OPINION   importance
readers         0.62     0.01     0.043214959104603226
opinion page    0.0      0.0      0.03793103448275861
website         1.27     0.04     0.033390786933015296
percent         2.88     0.07     0.03136646998033137
nexpand document 0.38     0.0      0.029310740541509773
report          3.43     0.47     0.024787115263662673
rumor           0.54     0.0      0.021739594433705964
board           0.52     0.89     0.019207823443225382
diverse opinions 0.0      0.46     0.016844513284880257
say             3.56     0.72     0.01660661644284311
d               202.41   92.46    0.015354847093476565
wrote           0.97     0.16     0.014960840852320741
text            0.66     0.23     0.014411211016855562
email           2.53     0.09     0.014038084861435074
m               163.73   67.84    0.011945141882609311
eligible        0.82     0.0      0.011828123192031876
cost            2.31     0.33     0.01172018700130058
angelo          0.29     0.01     0.011461463606678131
doesn           0.73     0.21     0.010132547864506625
care            5.36     0.44     0.010038167388167384
```

3) Unigrams and POS Tags

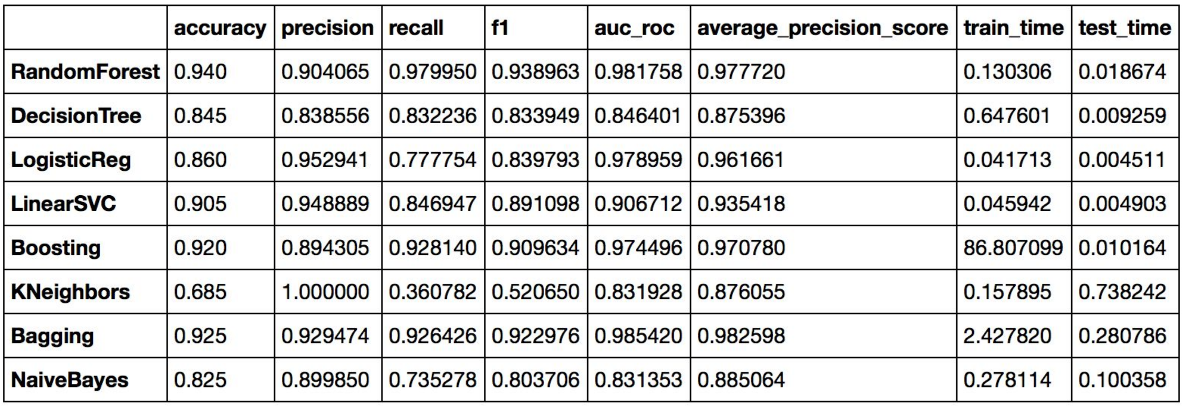


	accuracy	precision	recall	f1	auc_roc	average_precision_score	train_time	test_time
RandomForest	0.930	0.894622	0.974000	0.931214	0.987991	0.988756	0.028363	0.002705
DecisionTree	0.875	0.869827	0.886045	0.874321	0.879877	0.907936	0.075775	0.001093
LogisticReg	0.880	0.905195	0.874667	0.874780	0.977430	0.966179	0.008650	0.001199
LinearSVC	0.920	0.940909	0.888000	0.910607	0.921975	0.941955	0.009555	0.000789
Boosting	0.970	0.967380	0.974000	0.969387	0.997333	0.998499	11.348626	0.001804
KNeighbors	0.820	0.765635	0.918281	0.830655	0.880972	0.877867	0.030577	0.108627
Bagging	0.930	0.947895	0.919429	0.928040	0.974897	0.974445	0.363864	0.034470
NaiveBayes	0.750	0.846460	0.636421	0.709678	0.762135	0.836441	0.032398	0.011270

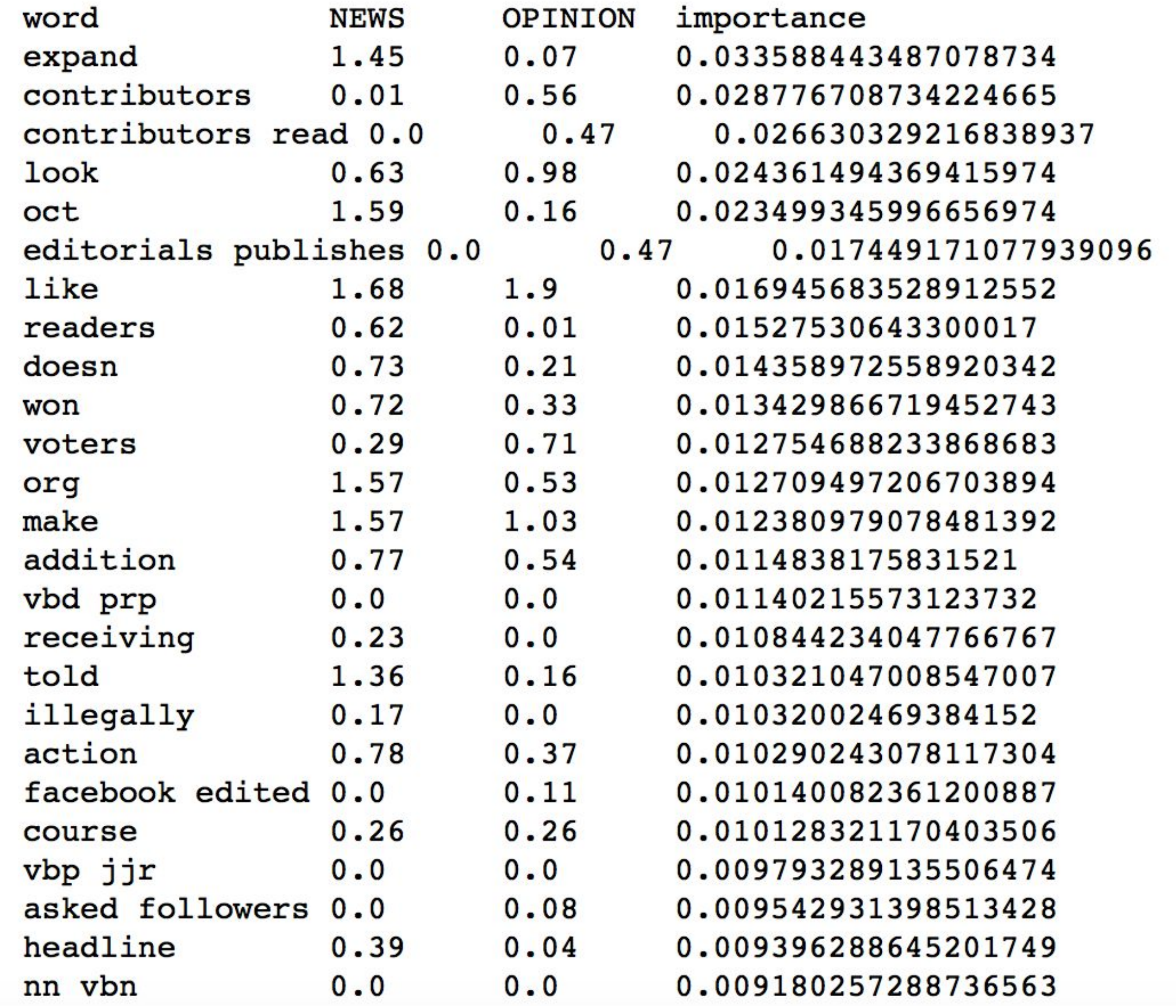


```
word              NEWS     OPINION  importance
department        1.75     0.25     0.05374790339023765
opinion           0.12     0.96     0.03838696683423041
like              1.07     1.5      0.03262028067526732
nexpand           0.38     0.0      0.02774369461486025
link              0.49     0.0      0.02365478210988829
number            1.32     0.13     0.02166624193317518
document          1.08     0.03     0.020571147279221817
factcheck         0.36     0.0      0.01965743892823152б
based             0.99     0.15     0.01961782880277227
didn              0.55     0.06     0.016094987081237865
true              1.9      0.14     0.01590407415438665
eugene            0.23     0.02     0.01576679265632012
published         0.39     0.09     0.015565477621122578
better            0.18     0.38     0.014689363336628136
publishes         0.01     0.47     0.01331693958954915б
percent           2.57     0.04     0.012951350481362487
right             0.62     0.71     0.012927221702813472
circulated        0.29     0.0      0.012837887535660595
claims            1.66     0.08     0.01216242569313486
included          0.36     0.03     0.011043373686774015
```

4) Unigrams, Bigrams and POS Tags



	accuracy	precision	recall	f1	auc_roc	average_precision_score	train_time	test_time
RandomForest	0.940	0.904065	0.979950	0.938963	0.981758	0.977720	0.130306	0.018674
DecisionTree	0.845	0.838556	0.832236	0.833949	0.846401	0.875396	0.647601	0.009259
LogisticReg	0.860	0.952941	0.777754	0.839793	0.978959	0.961661	0.041713	0.004511
LinearSVC	0.905	0.948889	0.846947	0.891098	0.906712	0.935418	0.045942	0.004903
Boosting	0.920	0.894305	0.928140	0.909634	0.974496	0.970780	86.807099	0.010164
KNeighbors	0.685	1.000000	0.360782	0.520650	0.831928	0.876055	0.157895	0.738242
Bagging	0.925	0.929474	0.926426	0.922976	0.985420	0.982598	2.427820	0.280786
NaiveBayes	0.825	0.899850	0.735278	0.803706	0.831353	0.885064	0.278114	0.100358



```
word            NEWS     OPINION  importance
expand          1.45     0.07     0.033588443487078734
contributors    0.01     0.56     0.028776708734224665
contributors read 0.0      0.47     0.026630329216838937
look            0.63     0.98     0.024361494369415974
oct             1.59     0.16     0.023499345996656974
editorials publishes 0.0      0.47     0.017449171077939096
like            1.68     1.9      0.016945683528912552
readers         0.62     0.01     0.015275306433000017
doesn           0.73     0.21     0.014358972558920342
won             0.72     0.33     0.013429866719452743
voters          0.29     0.71     0.012754688233868683
org             1.57     0.53     0.012709497206703894
make            1.57     1.03     0.012380979078481392
addition        0.77     0.54     0.0114838175831521
vbd prp         0.0      0.0      0.011402155731237
receiving       0.23     0.0      0.010844234047766767
told            1.36     0.16     0.010321047008547007
illegally       0.17     0.0      0.01032002469384152
action          0.78     0.37     0.010290243078117304
facebook edited 0.0      0.11     0.010140082361200887
course          0.26     0.26     0.010128321170403506
vbp jjr         0.0      0.0      0.009793289135506474
asked followers 0.0      0.08     0.009542931398513428
headline        0.39     0.04     0.009396288645201749
nn vbn          0.0      0.0      0.009180257288736563
```