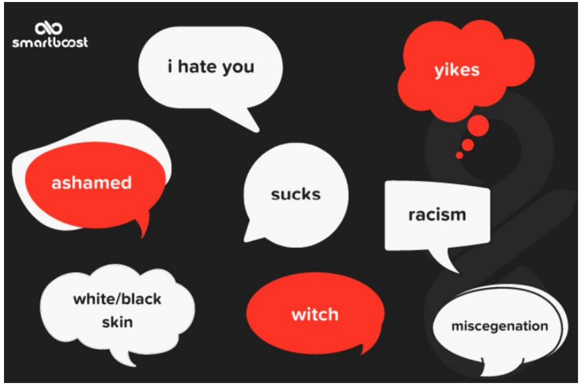
CIS 5300 Project

Noam Elul, Dan Gallagher, Aman Kumar, Yuling Liu

12/22/2022

Cross-Domain Detection of Hate-Speech

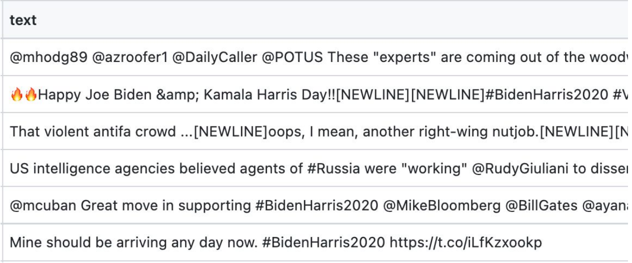
- "Don't judge me.. i am vindictive and vengeful when it comes to **bitches** disrespecting me" - Twitter Example
- "It should definitely say something about Kyle Vander Wielen and **all his bitchin**" - Wikipedia Example
- "Public speech that expresses **hate** or encourages **violence towards a person or group** based on something such as race, religion, sex, or sexual orientation" - Cambridge Dictionary



Why Cross-Domain Hate-Speech Detection

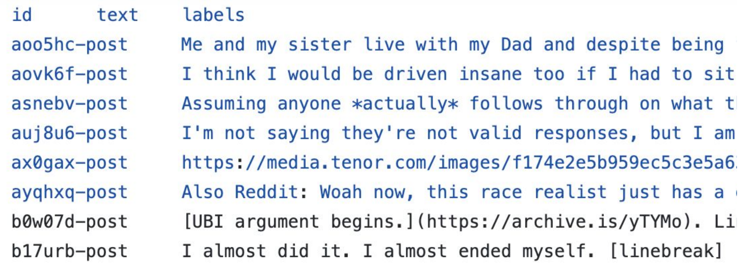
- Hate speech is prevalent on social media sites
- There exist many labeled datasets, but it’s difficult to generate datasets for every different website
- So we decided to **train a toxicity classifier using a dataset from one website, and measure its performance on a dataset from another website**
- Might generalize well to websites where we don’t have data
    - New social networks

Data



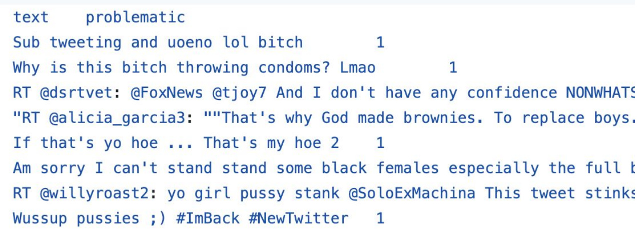
text
@mhodg89 @azroofer1 @DailyCaller @POTUS These "experts" are coming out of the wood
🔥🔥Happy Joe Biden &amp; Kamala Harris Day!![NEWLINE][NEWLINE]#BidenHarris2020 #
That violent antifa crowd ...[NEWLINE]oops, I mean, another right-wing nutjob.[NEWLINE][
US intelligence agencies believed agents of #Russia were "working" @RudyGiuliani to disse
@mcuban Great move in supporting #BidenHarris2020 @MikeBloomberg @BillGates @ayan
Mine should be arriving any day now. #BidenHarris2020 https://t.co/iLfKzxookp

Political Twitter Hate Speech



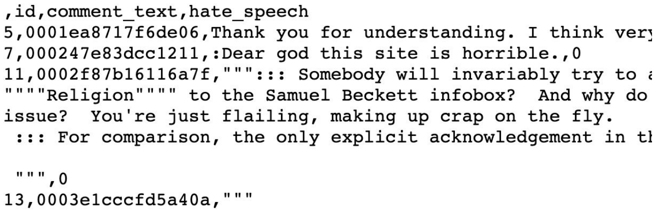
```
id       text      labels
aoo5hc-post        Me and my sister live with my Dad and despite being
aovk6f-post        I think I would be driven insane too if I had to sit
asnebv-post        Assuming anyone *actually* follows through on what t
auj8u6-post        I'm not saying they're not valid responses, but I am
ax0gax-post        https://media.tenor.com/images/f174e2e5b959ec5c3e5a6
ayqhxq-post        Also Reddit: Woah now, this race realist just has a
b0w07d-post        [UBI argument begins.](https://archive.is/yTYMo). Li
b17urb-post        I almost did it. I almost ended myself. [linebreak]
```

Reddit Abusive Comments



```
text    problematic
Sub tweeting and uoeno lol bitch        1
Why is this bitch throwing condoms? Lmao        1
RT @dsrtvet: @FoxNews @tjoy7 And I don't have any confidence NONWHAT
"RT @alicia_garcia3: ""That's why God made brownies. To replace boys.
If that's yo hoe ... That's my hoe 2    1
Am sorry I can't stand stand some black females especially the full
RT @willyroast2: yo girl pussy stank @SoloExMachina This tweet stink
Wussup pussies ;) #ImBack #NewTwitter   1
```

Twitter Hate Speech



```
,id,comment_text,hate_speech
5,0001ea8717f6de06,Thank you for understanding. I think ver
7,000247e83dcc1211,:Dear god this site is horrible.,0
11,0002f87b16116a7f,""":: Somebody will invariably try to
""""Religion"""" to the Samuel Beckett infobox?  And why do
issue?  You're just flailing, making up crap on the fly.
 ::: For comparison, the only explicit acknowledgement in t

 """,0
13,0003e1cccfd5a40a,"""
```

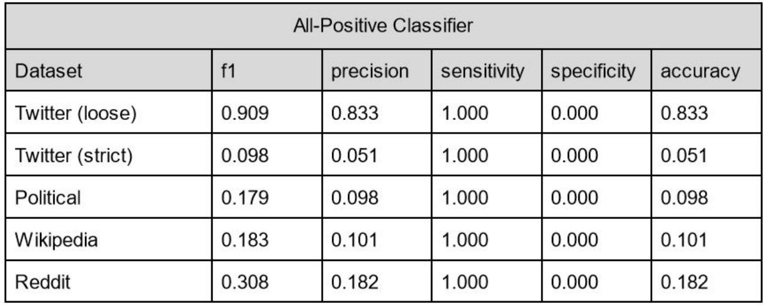
Wikipedia Hate Speech

Evaluation Metrics

- **F1 score (main)**
  - The harmonic mean of precision and sensitivity (recall for documents being classified as problematic)
- **Precision & Recall (Sensitivity and Specificity)**
  - Shows how a model handles class imbalance, where a model that completely ignores the minority class will still achieve high accuracy
- **Accuracy**
  - Helps get a sense of which ways a model might be succeeding, but it is important to note that a high accuracy in the absence of high performance on the other metrics does not signify a successful model

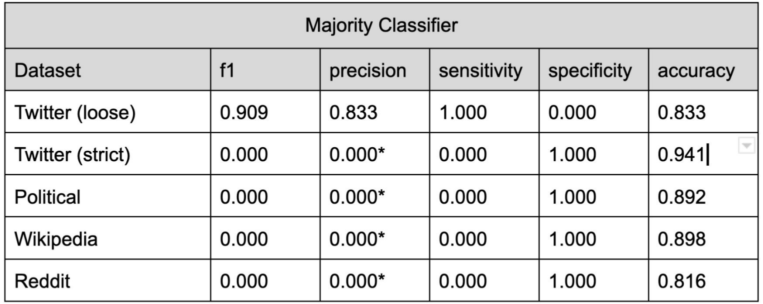
Simple Baselines

- All-positive Classifier



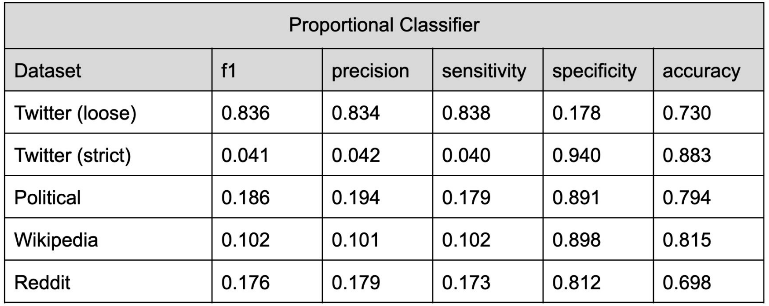
All-Positive Classifier					
Dataset	f1	precision	sensitivity	specificity	accuracy
Twitter (loose)	0.909	0.833	1.000	0.000	0.833
Twitter (strict)	0.098	0.051	1.000	0.000	0.051
Political	0.179	0.098	1.000	0.000	0.098
Wikipedia	0.183	0.101	1.000	0.000	0.101
Reddit	0.308	0.182	1.000	0.000	0.182

- Majority Classifier



Majority Classifier					
Dataset	f1	precision	sensitivity	specificity	accuracy
Twitter (loose)	0.909	0.833	1.000	0.000	0.833
Twitter (strict)	0.000	0.000*	0.000	1.000	0.941
Political	0.000	0.000*	0.000	1.000	0.892
Wikipedia	0.000	0.000*	0.000	1.000	0.898
Reddit	0.000	0.000*	0.000	1.000	0.816

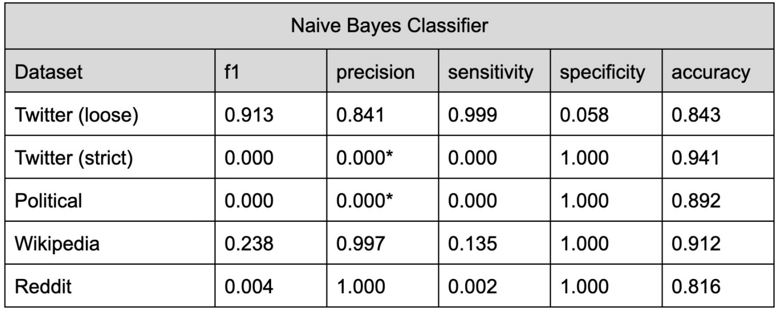
- Proportional Classifier



Proportional Classifier					
Dataset	f1	precision	sensitivity	specificity	accuracy
Twitter (loose)	0.836	0.834	0.838	0.178	0.730
Twitter (strict)	0.041	0.042	0.040	0.940	0.883
Political	0.186	0.194	0.179	0.891	0.794
Wikipedia	0.102	0.101	0.102	0.898	0.815
Reddit	0.176	0.179	0.173	0.812	0.698

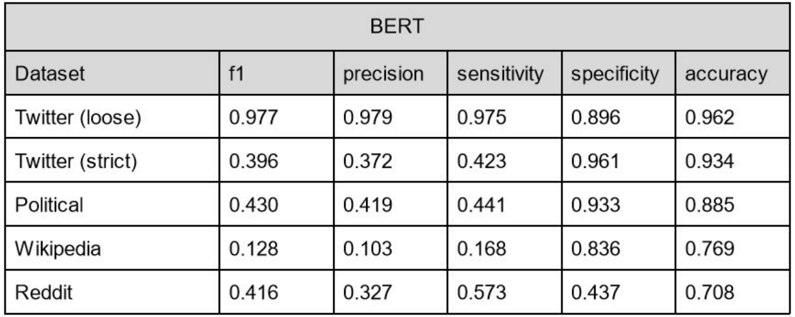
Strong Baselines

- ## Naive Bayes on TF-IDF Representation



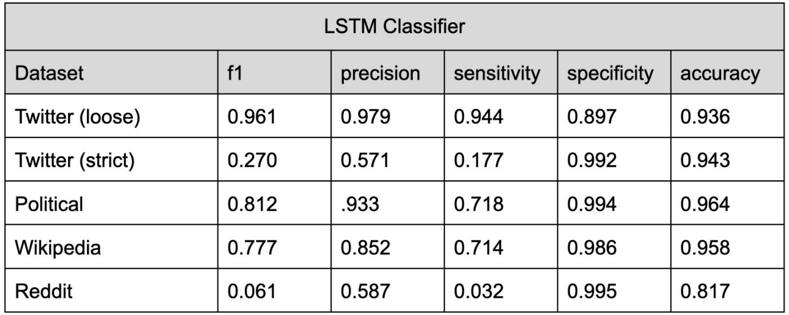
Naive Bayes Classifier					
Dataset	f1	precision	sensitivity	specificity	accuracy
Twitter (loose)	0.913	0.841	0.999	0.058	0.843
Twitter (strict)	0.000	0.000*	0.000	1.000	0.941
Political	0.000	0.000*	0.000	1.000	0.892
Wikipedia	0.238	0.997	0.135	1.000	0.912
Reddit	0.004	1.000	0.002	1.000	0.816

- ## BERT



BERT					
Dataset	f1	precision	sensitivity	specificity	accuracy
Twitter (loose)	0.977	0.979	0.975	0.896	0.962
Twitter (strict)	0.396	0.372	0.423	0.961	0.934
Political	0.430	0.419	0.441	0.933	0.885
Wikipedia	0.128	0.103	0.168	0.836	0.769
Reddit	0.416	0.327	0.573	0.437	0.708

- ## LSTM

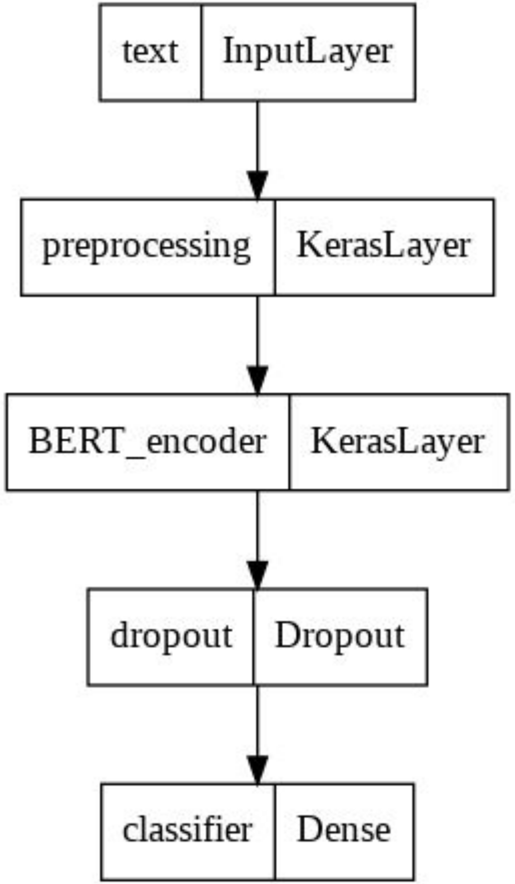


LSTM Classifier					
Dataset	f1	precision	sensitivity	specificity	accuracy
Twitter (loose)	0.961	0.979	0.944	0.897	0.936
Twitter (strict)	0.270	0.571	0.177	0.992	0.943
Political	0.812	.933	0.718	0.994	0.964
Wikipedia	0.777	0.852	0.714	0.986	0.958
Reddit	0.061	0.587	0.032	0.995	0.817

BERT Architecture

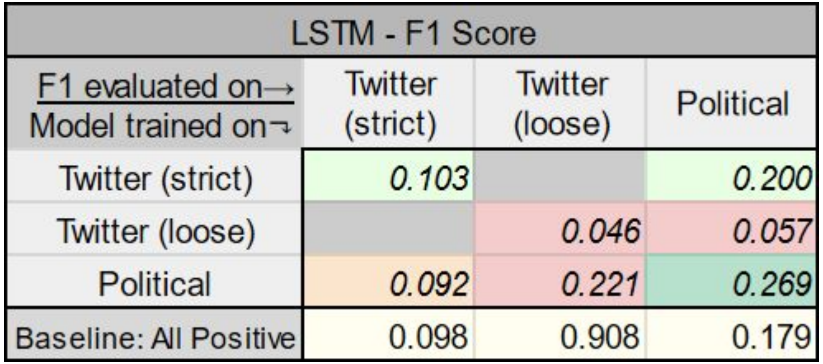
$$TP = \sum y_{true} \cdot y_{pred}$$
$$TN = \sum (1 - y_{true}) \cdot (1 - y_{pred})$$
$$FP = \sum (1 - y_{true}) \cdot y_{pred}$$
$$FN = \sum y_{true} \cdot (1 - y_{pred})$$

- We used a pretrained BERT model from the Keras library as another strong classifier
  - We used a small BERT model with 4 hidden layers, a hidden size of 512, and 8 attention heads, to keep training times manageable. Larger BERTs did not seem to improve performance.
- We experimented with different loss functions
  - Cross-Entropy Loss: Due to the class imbalance, did not generally produce good results.
  - Macro F1 Loss: A loss function based on 1 minus a modified F1 score, where true/false positives/negatives are calculated as the sum of the predicted probabilities. Generally worked better, but sometimes caused the model to get “stuck” only predicting positives.



Extension 1:
Same-Platform Transferability

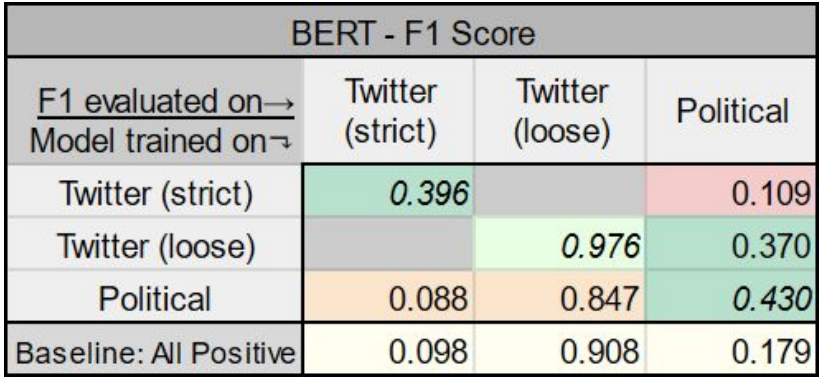
- Across the same platform (Twitter), the BERT model showed promise in transferring between Twitter (Loose) and Political
  - The f1 score when evaluating on Political was low, but not much lower than when trained on Political
  - In both directions, the f1 score of the transferred model was ~86% of the score of the original model
- BERT transferred significantly better than LSTM
- Twitter (Strict) did not transfer well
  - Annotation scheme may be too different from others
  - Problem of differentiating hate-speech from offensive speech is more difficult than differentiating offensive speech from non-offensive speech



LSTM - F1 Score

F1 evaluated on→ Model trained on¬	Twitter (strict)	Twitter (loose)	Political
Twitter (strict)	*0.103*		*0.200*
Twitter (loose)		*0.046*	*0.057*
Political	*0.092*	*0.221*	*0.269*
Baseline: All Positive	0.098	0.908	0.179

Colors indicate performance relative to baseline

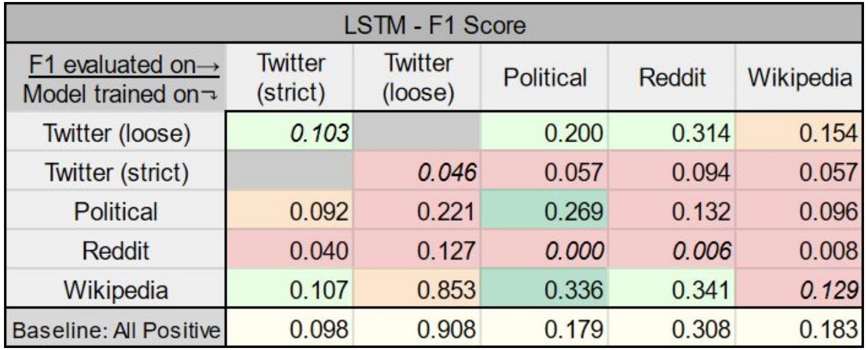


BERT - F1 Score

F1 evaluated on→ Model trained on¬	Twitter (strict)	Twitter (loose)	Political
Twitter (strict)	*0.396*		0.109
Twitter (loose)		*0.976*	0.370
Political	0.088	0.847	*0.430*
Baseline: All Positive	0.098	0.908	0.179

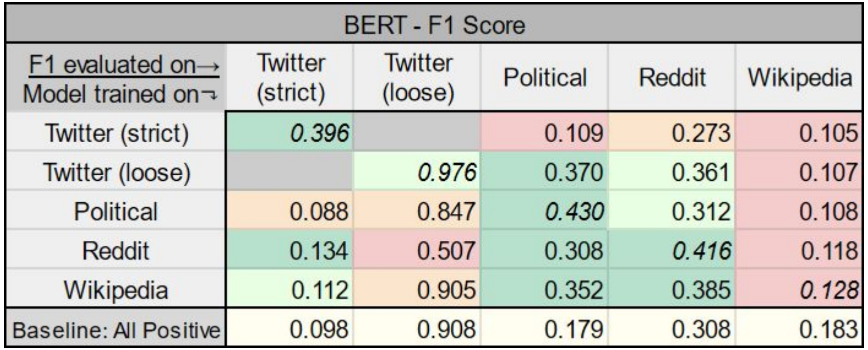
Extension 2:
Cross-Platform Transferability

- When transferring across different websites, performance was not great, but wasn't much worse than across the same website.
  - While the performance of the Wikipedia dataset did not degrade much when trained on other datasets compared to when trained on itself, it's hard to conclude how well the transfer worked, since the F1 was already so low that it may simply be hitting a floor since performance can't get much worse.
  - On the other hand, when *training* using the Wikipedia dataset, transfer performance was generally fairly good.
  - The performance of the Reddit dataset did not degrade much when training on the Twitter (loose) or Wikipedia datasets.



LSTM - F1 Score					
F1 evaluated on→ Model trained on¬	Twitter (strict)	Twitter (loose)	Political	Reddit	Wikipedia
Twitter (loose)	*0.103*		0.200	0.314	0.154
Twitter (strict)		*0.046*	0.057	0.094	0.057
Political	0.092	0.221	0.269	0.132	0.096
Reddit	0.040	0.127	*0.000*	*0.006*	0.008
Wikipedia	0.107	0.853	0.336	0.341	*0.129*
Baseline: All Positive	0.098	0.908	0.179	0.308	0.183

Colors indicate performance relative to baseline



BERT - F1 Score					
F1 evaluated on→ Model trained on¬	Twitter (strict)	Twitter (loose)	Political	Reddit	Wikipedia
Twitter (strict)	*0.396*		0.109	0.273	0.105
Twitter (loose)		*0.976*	0.370	0.361	0.107
Political	0.088	0.847	*0.430*	0.312	0.108
Reddit	0.134	0.507	0.308	*0.416*	0.118
Wikipedia	0.112	0.905	0.352	0.385	*0.128*
Baseline: All Positive	0.098	0.908	0.179	0.308	0.183

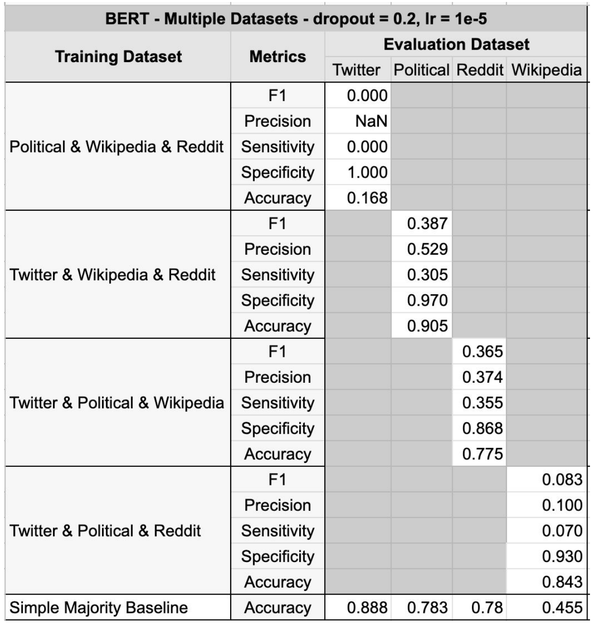
Further Extension: Multiple Dataset

3 datasets combined

- Low F1, low precision, and low recall
- Only improved on the Political dataset
- Multiple contexts may be useful sometimes, but not always

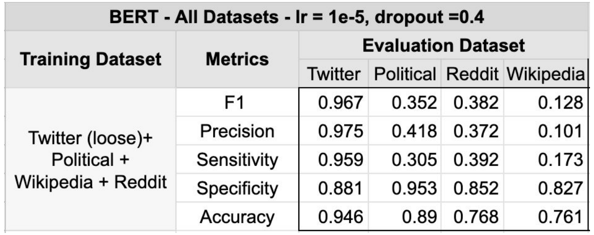
4 datasets combined

- Performed well on twitter dataset
- Performed well on the reddit dataset relative to our previous models in terms of F1 score



BERT - Multiple Datasets - dropout = 0.2, lr = 1e-5

Training Dataset	Metrics	Evaluation Dataset			
		Twitter	Political	Reddit	Wikipedia
Political & Wikipedia & Reddit	F1	0.000			
	Precision	NaN			
	Sensitivity	0.000			
	Specificity	1.000			
	Accuracy	0.168			
Twitter & Wikipedia & Reddit	F1		0.387		
	Precision		0.529		
	Sensitivity		0.305		
	Specificity		0.970		
	Accuracy		0.905		
Twitter & Political & Wikipedia	F1			0.365	
	Precision			0.374	
	Sensitivity			0.355	
	Specificity			0.868	
	Accuracy			0.775	
Twitter & Political & Reddit	F1				0.083
	Precision				0.100
	Sensitivity				0.070
	Specificity				0.930
	Accuracy				0.843
Simple Majority Baseline	Accuracy	0.888	0.783	0.78	0.455



BERT - All Datasets - lr = 1e-5, dropout =0.4

Training Dataset	Metrics	Evaluation Dataset			
		Twitter	Political	Reddit	Wikipedia
Twitter (loose)+ Political + Wikipedia + Reddit	F1	0.967	0.352	0.382	0.128
	Precision	0.975	0.418	0.372	0.101
	Sensitivity	0.959	0.305	0.392	0.173
	Specificity	0.881	0.953	0.852	0.827
	Accuracy	0.946	0.89	0.768	0.761

Conclusion

- Overall, there is some promise in BERT models being able to transfer across datasets of both different structure and content
- BERT outperformed LSTM in both individual performance on each dataset as well as the ability to transfer between datasets
- Other than on general tweets, our models did not effectively handle the massive class imbalance present. This makes our transferability results a bit less conclusive

What we learned

- We explored ways to combat class imbalance in text
  - Undersampling
  - Test augmentation
  - Focal loss
- We applied BERT
  - Using the architecture from Keras
- We explored how models transfer across datasets
  - Cross-domain application of a model is not perfect, but may hold promise as a stopgap to detect hate-speech in burgeoning social networks to allow for data to be properly annotated