TWITTER HATE SPEECH ANALYSIS

MILESTONE TWO

COLIN GREEN & SEAN ZHANG

FEATURE CREATION AND EXTRACTION

- We created the following features:
  - num_tokens, num_mentions, num_urls, num_hashtags
- We tested out the following features:
  - Bag of words, ngrams (n=2, & n=3), TF-IDF
  - Word2Vec - CBOW, concat(Word2Vec - CBOW, BOW), and Custom - CBOW
- We ran logistic regression and decision tree on each embedding as a benchmark

DATA BEFORE EMBEDDING

class	tweet	tweet_clean	tweet_lemma	num_tokens	tweet_no_others	mention_count	url_count	hashtag_count
1	" bitch who do you love "	bitch love	bitch love	2.0	bitch love	0	0	0
1	" fuck no that bitch dont even suck dick " &#1...	fuck bitch dont even suck dick kermit videos b...	fuck bitch do not even suck dick ...	13.0	fuck bitch dont even suck dick kermit videos b...	0	0	0
1	" lames crying over hoes thats tears of a clown "	lames crying hoes thats tears clown	lame cry hoe that s tear clown	7.0	lames crying hoes thats tears clown	0	0	0
1	"..All I wanna do is get money and fuck model ...	all i wanna get money fuck model bitches russe...	all i wanna get money fuck model bitch ru...	11.0	all i wanna get money fuck model bitches russe...	0	0	0
1	"@ARIZZLEINDACUT: Females think dating a pussy...	females think dating pussy cute now stuff make...	mentionhere female think date pussy cute now...	14.0	females think dating pussy cute now stuff make...	1	1	0

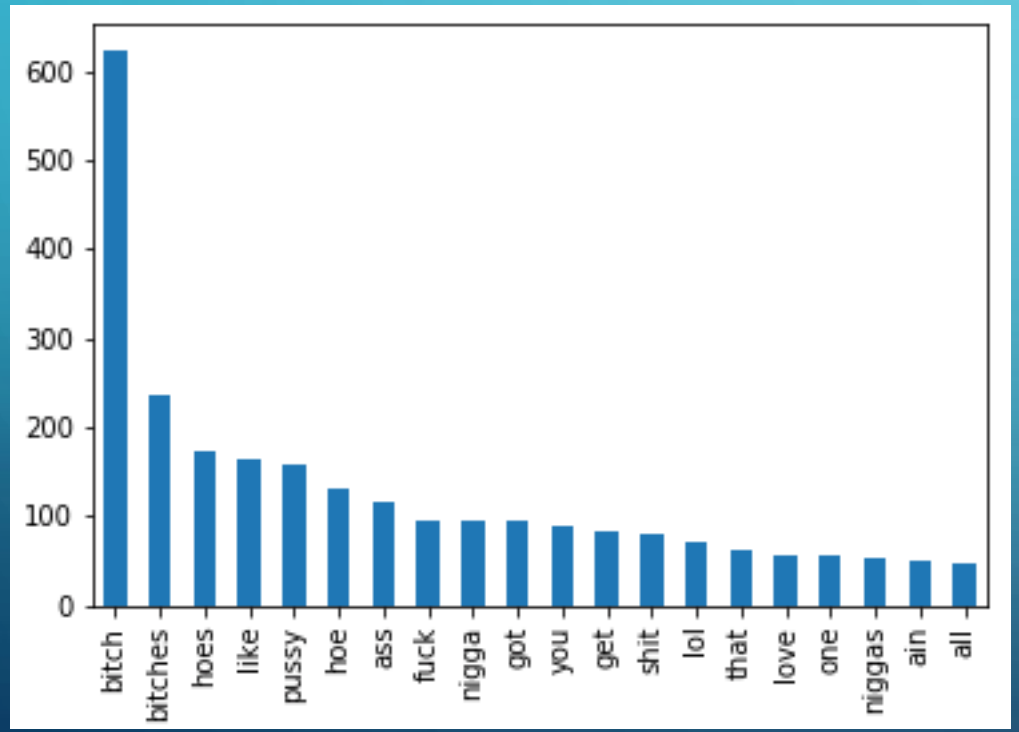
BAG OF WORDS

	aa	aaaaaaaaand	aap	aaron	aaronmacgruder	ab	ability	abortion	about	abraham	...	zimmerman	zimmy	zion	zionist	zipperheads	zoe	zog	zo
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2855	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
2856	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
2857	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
2858	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	
2859	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	

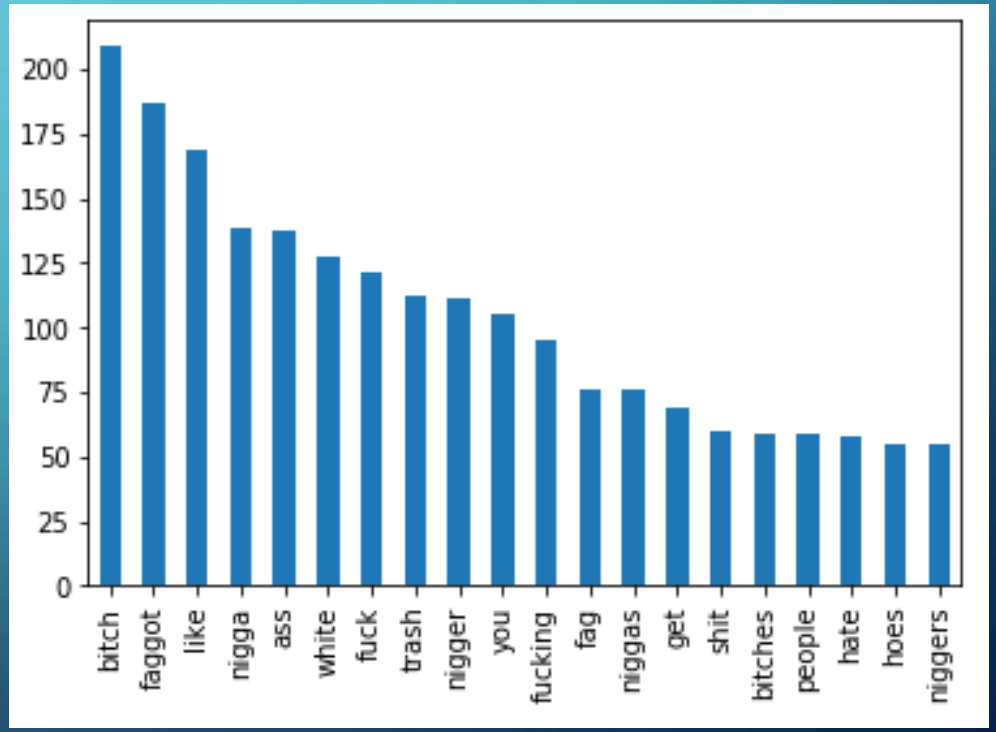
2860 rows × 5139 columns

BAG OF WORDS – FEATURE COUNTS

Offensive Tweets



Hate Tweets

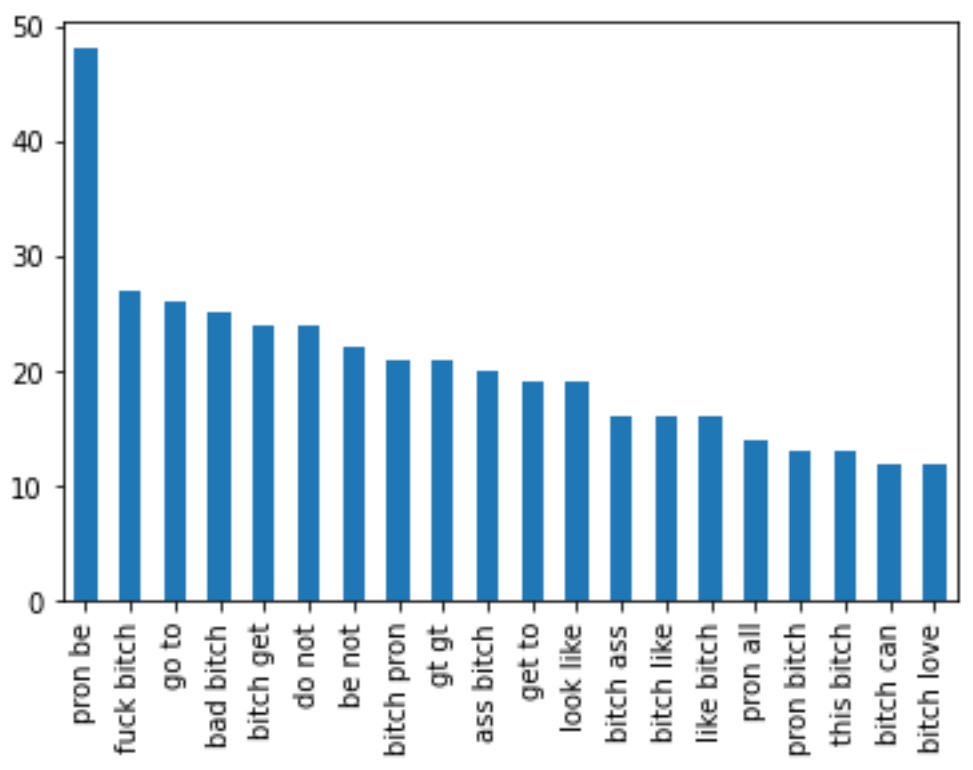


NGRAMS (N=2)

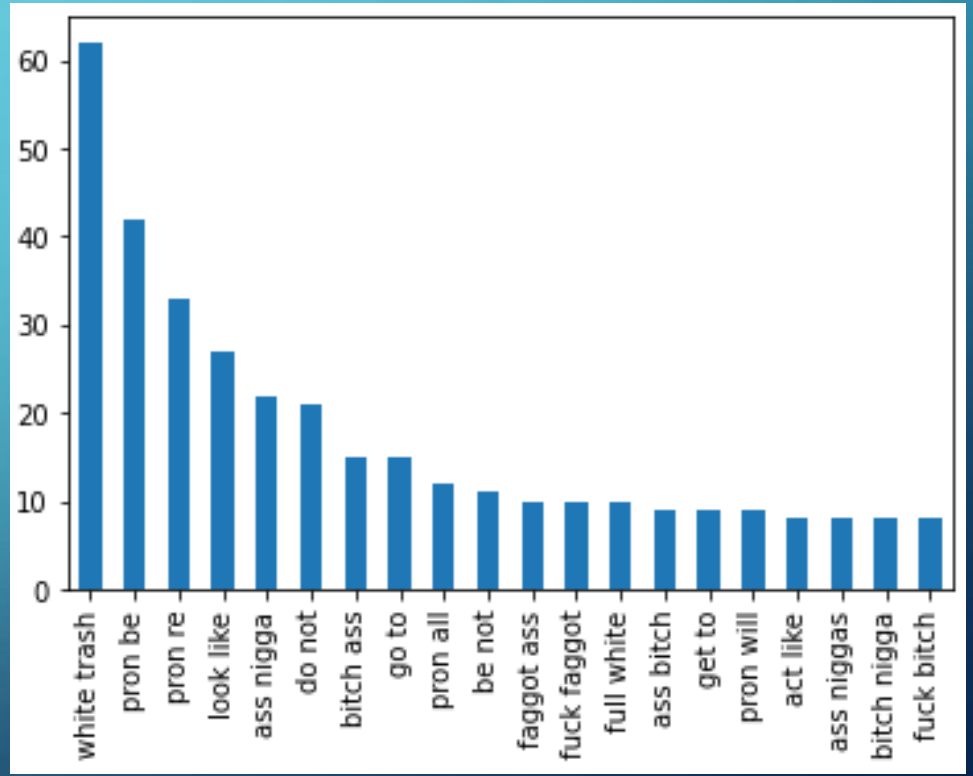
aronmacgruder stuff	ability block	abortion get	about act	abraham lincoln	absolute pussy	...	zimmerman arrest	zimmerman comin	zimmerman creepy	zimmerman jury	zimmy attack	zionist codeword	zoe ask	zog jews	zone radius	zoom see
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

NGRAMS (N=2) – FEATURE COUNT

Offensive Tweets



Hate Tweets

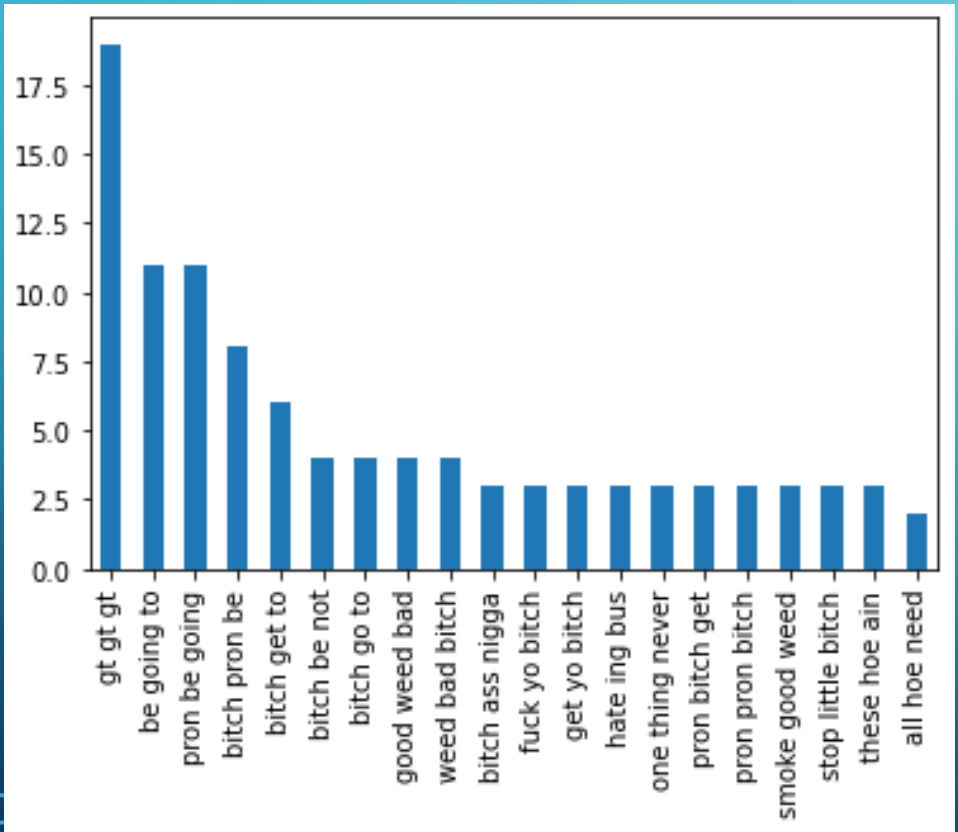


NGRAMS (N=3)

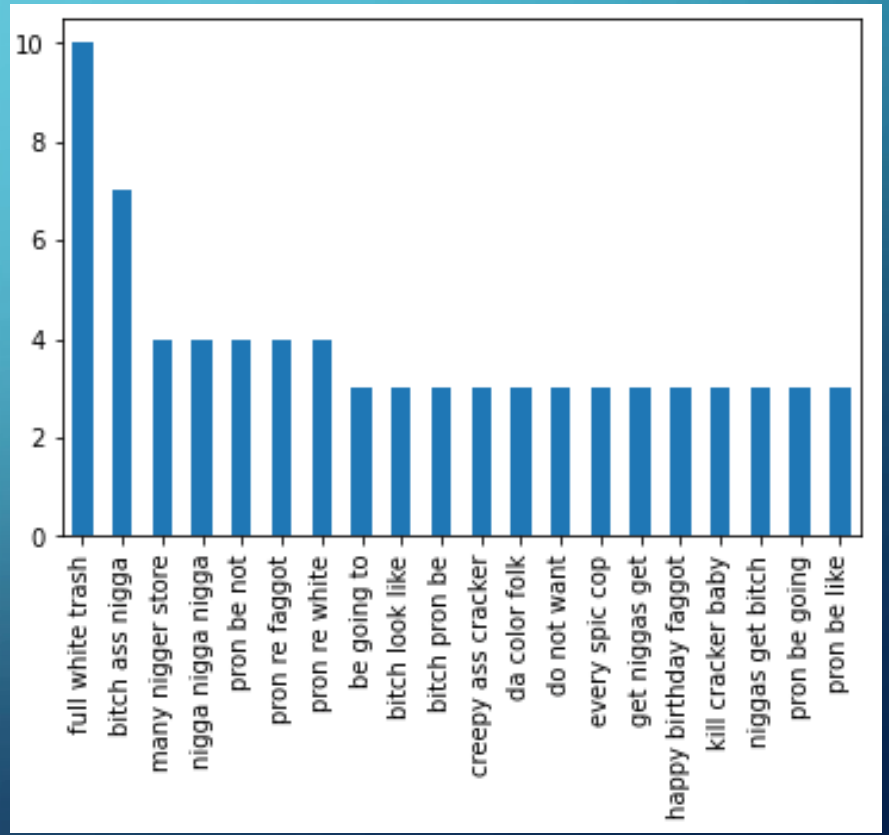
abortion get cemetery	about act color	abraham lincoln quote	absolute pussy scared	absolve end today	abt bitch face	...	zimmerman arrest do	zimmerman comin yo	zimmerman creepy ass	zimmerman jury ggas	zimmy attack traytray	zionist codeword antisemite	zoe ask account	zog jews run	zone radius around	zoom see hairy
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

NGRAMS (N=3) – FEATURE COUNT

Offensive Tweets



Hate Tweets



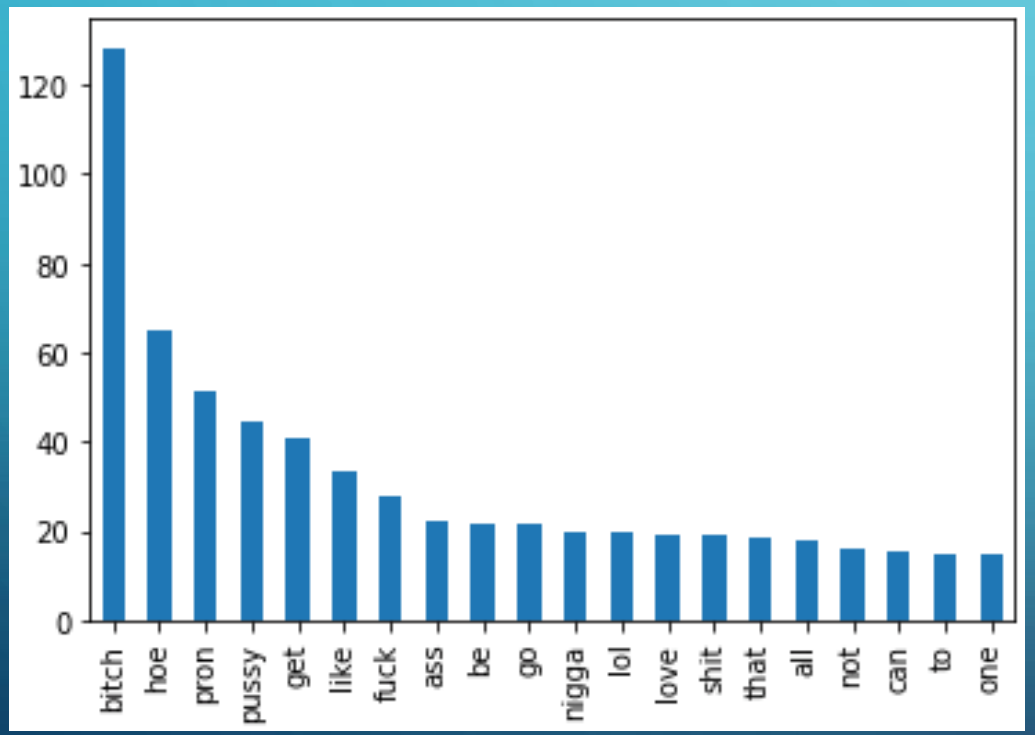
TF-IDF

	aa	aaaaaaaaand	aap	aaron	aaronmacgruder	ab	ability	abortion	about	abraham	...	zimmerman	zimmy	zion	zionist	zipperhead	zoe	zog	zor
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2855	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2856	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2857	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2858	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2859	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

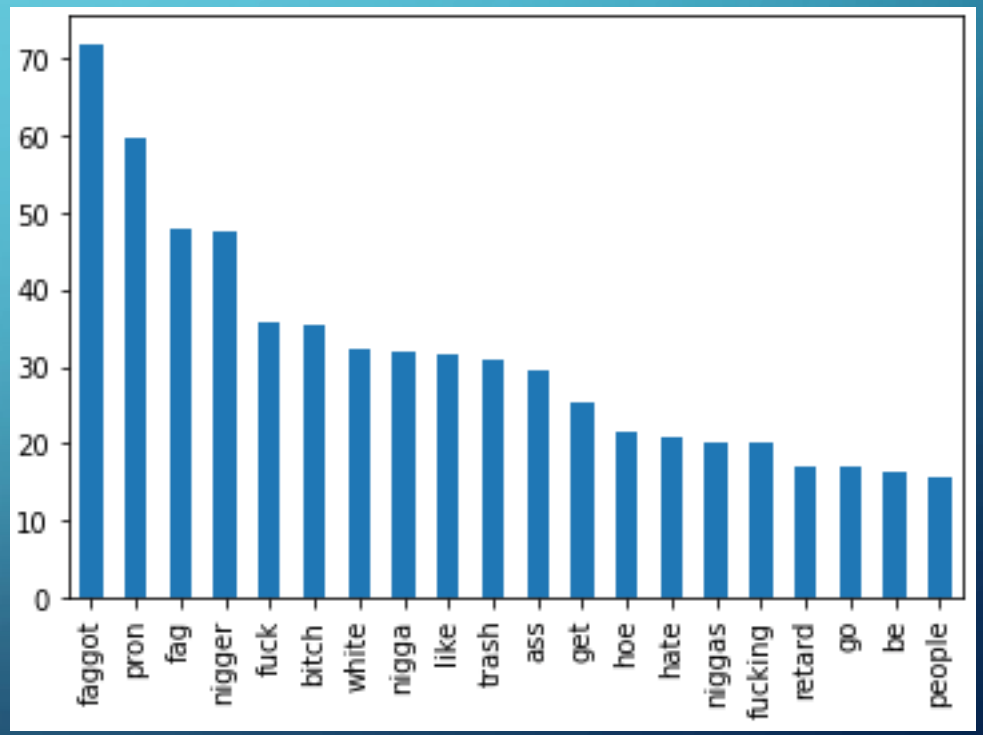
2860 rows × 4451 columns

TF-IDF – FEATURE COUNT

Offensive Tweets



Hate Tweets



WORD2VEC - CBOW

	0	1	2	3	4	5	6	7	8	9	...	90	91	92	93
0	0.426253	-0.449129	-0.306192	-0.512622	-0.051190	-0.849438	-0.727272	-0.114577	0.209313	0.542770	...	-0.536479	-0.510480	-0.453063	-0.974342
1	0.153138	-0.481615	-0.220396	0.093086	-0.078921	-0.610603	-0.291115	0.172201	0.141143	0.147177	...	-0.201854	-0.239412	-0.111255	-0.650069
2	0.122567	-0.151198	-0.117174	-0.185195	0.031289	-0.524518	-0.244605	-0.147231	0.130114	0.206408	...	-0.109597	-0.119751	-0.008681	-0.375946
3	0.416339	-0.514383	-0.317563	-0.589974	0.079081	-0.780523	-0.554621	-0.077496	0.176782	0.352400	...	-0.327384	-0.452492	-0.572076	-0.720230
4	0.369659	-0.333850	-0.160261	-0.234247	-0.306322	-0.424265	-0.175677	-0.147177	0.219274	0.377423	...	0.044297	0.012140	0.114466	-0.160522
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2855	0.443163	-0.569709	-0.311475	-0.180331	-0.698957	-0.226150	-0.245503	0.148401	-0.254065	0.354032	...	-0.249894	-0.558521	-0.262517	-0.361339
2856	1.144400	-0.726996	-0.211574	-0.924790	-0.633785	-0.134006	-0.575602	-0.547512	-0.694805	0.512575	...	1.237443	-0.711082	-0.077753	-0.215808
2857	0.337430	-0.429522	-0.241379	0.051542	-0.289824	-0.051472	-0.058383	-0.006023	-0.200190	-0.024186	...	0.151084	-0.291540	0.140829	-0.267152
2858	0.240839	-0.029532	-0.174792	-0.257355	-0.399906	-0.690719	-0.261135	0.408957	0.322165	0.576915	...	-0.152155	-0.341586	-0.382100	-0.980751
2859	0.187630	-0.122916	-0.123539	-0.348902	0.064622	-0.488635	-0.274453	-0.127944	0.260815	0.355561	...	-0.273105	-0.147576	-0.197073	-0.490919

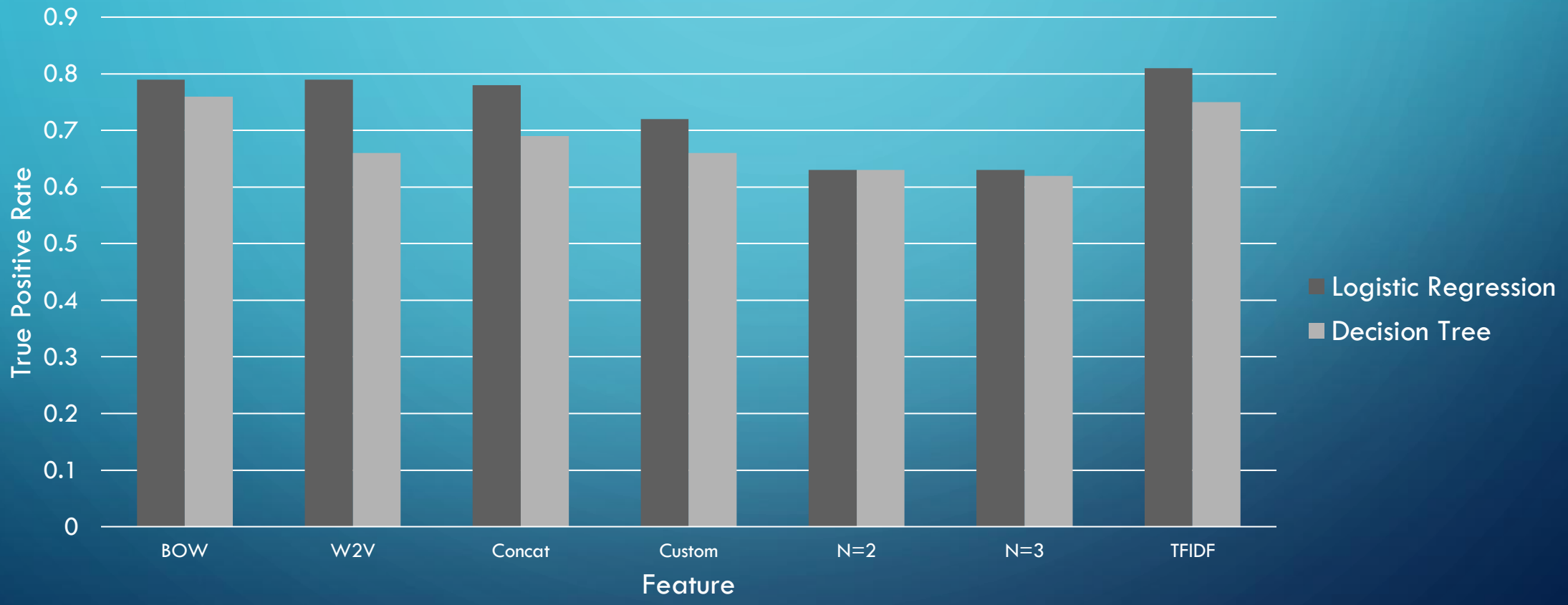
2860 rows × 100 columns

CUSTOM - CBOW

	0	1	2	3	4	5	6	7	8	9	...	
0	-0.092240	0.655190	-0.054113	-0.001526	0.245873	-0.047122	-0.195120	-0.241070	-0.024567	-0.028244	...	-0.009(
1	-0.036190	0.052367	0.056565	0.040076	0.153519	-0.078184	-0.050946	-0.083378	-0.090013	0.050202	...	0.086
2	-0.084097	0.030291	0.036536	0.039797	0.106109	0.002597	0.060696	-0.047483	-0.047069	0.039118	...	0.095
3	0.007605	0.010492	0.032044	0.126457	0.047635	-0.069211	-0.054946	-0.011408	-0.095851	0.020046	...	0.103(
4	-0.100222	-0.060809	0.076636	-0.023275	0.134729	-0.025185	0.020398	-0.055950	-0.117073	-0.033665	...	0.097
...	...	...	...	...	...	...	...	...	...	...	...	
2855	0.082106	0.163869	0.087209	0.145215	0.068083	-0.082497	0.035079	0.005989	-0.094450	0.008168	...	0.123
2856	-0.178984	0.105335	0.093642	0.143920	0.036653	-0.008155	-0.112179	-0.071968	-0.040059	0.151145	...	0.115(
2857	-0.014508	0.042220	0.106782	0.091900	0.062218	-0.030795	-0.054249	-0.053927	-0.074365	0.050654	...	0.079
2858	0.001350	0.051687	0.038868	0.093754	0.076561	-0.059643	0.140547	-0.080700	-0.093650	0.042169	...	0.129
2859	0.031242	0.092824	0.032311	0.049737	0.053893	-0.075962	-0.002365	0.022531	-0.050144	0.002471	...	0.066

2860 rows × 100 columns

FEATURE SELECTION – TRUE POSITIVE RATE



BEST MODEL – LOGISTIC REGRESSION ON TF-IDF

	precision	recall	f1-score	support
0	0.79	0.79	0.79	341
1	0.81	0.81	0.81	374
accuracy			0.80	715
macro avg	0.80	0.80	0.80	715
weighted avg	0.80	0.80	0.80	715

FEATURE IMPORTANCE
TF-IDF – LOGISTIC REGRESSION

	features	importance
1253	faggot	-4.594443
2639	nigger	-3.856314
2636	niggas	-2.768235
2633	nigga	-2.725411
4290	white	-2.599723
1251	fag	-2.595279
3097	queer	-1.813521
1493	gay	-1.792469
793	coon	-1.697835
2096	kill	-1.667324
2870	people	-1.607309
1679	hate	-1.598636
4014	trash	-1.503182
400	black	-1.423769
3213	retard	-1.385454
1430	fuck	-1.372237
317	beaner	-1.347287
4269	wetback	-1.224424
3539	smh	-1.216035
3110	racist	-1.166534

FEATURE IMPORTANCE
TF-IDF – DECISION TREE

	features	importance			
388	bitch	0.158461	**4290**	white	0.008878
1759	hoe	0.102490	**1430**	fuck	0.008862
3087	pussy	0.085965	**1544**	go	0.008187
2633	nigga	0.034525	**89**	all	0.007932
2636	niggas	0.024863	**3433**	shit	0.007932
4450	num_tokens	0.020830	**4096**	ugly	0.006483
204	ass	0.012480	**4451**	mention_count	0.006480
2635	niggah	0.012407	**4065**	twat	0.006473
874	cunt	0.009711	**1253**	faggot	0.006466
4136	ur	0.009415	**390**	bitches	0.006407

SIMILAR WORDS (FROM FEATURE IMPORTANCE)

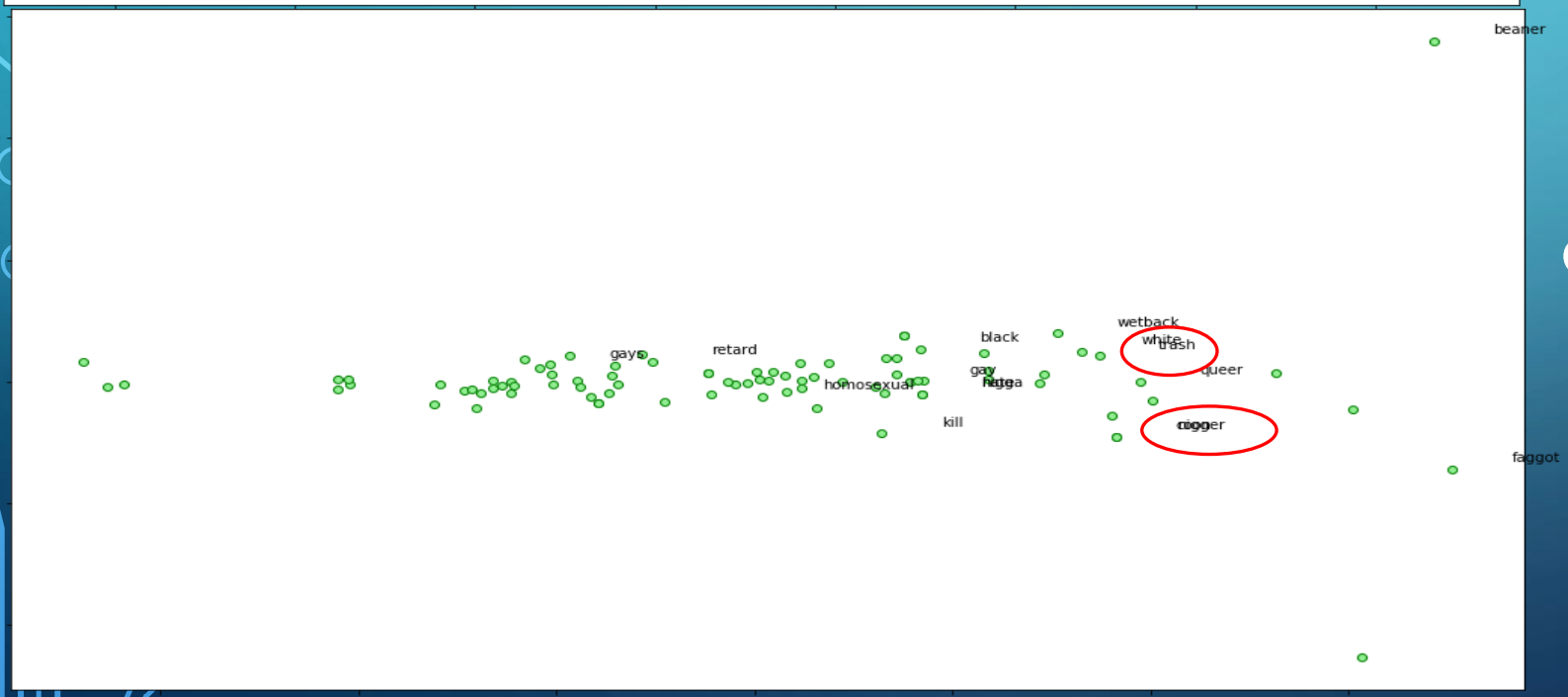
- WORD2VEC

```
{'faggot': ['tear', 'fag', 'sissy', 'powered', 'overly'],
 'nigger': ['hoodrats', 'traditions', 'tyler', 'honor', 'relevant'],
 'nigga': ['lame', 'scary', 'twin', 'dont', 'tryna'],
 'white': ['trash', 'westbrook', 'ducked', 'texarkana', 'slave'],
 'queer': ['pathetic', 'project', 'obama', 'yost', 'gaywrites'],
 'gay': ['welcome', 'episode', 'sucks', 'schedule', 'lord'],
 'coon': ['ban', 'tweets', 'ratchet', 'shoulda', 'its'],
 'kill': ['cops', 'panthers', 'presser', 'hijack', 'they'],
 'hate': ['dairy', 'ing', 'cripples', 'goddamit', 'bus'],
 'trash': ['white', 'westbrook', 'cocaine', 'portsmouth', 'slave'],
 'black': ['faves', 'pack', 'trailer', 'tbb', 'the'],
 'retard': ['guinea', 'claiming', 'republican', 'span', 'anyone'],
 'beaner': ['pussyed', 'homeless', 'pretend', 'lizard', 'opinion'],
 'wetback': ['lb', 'muslim', 'lethal', 'mike', 'weapon'],
 'homosexual': ['infiltration', 'priesthood', 'gear', 'voiced', 'eyebrow'],
 'gays': ['wholesome', 'ag', 'humble', 'takin', 'perm']}
```

- CUSTOM-TRAINED

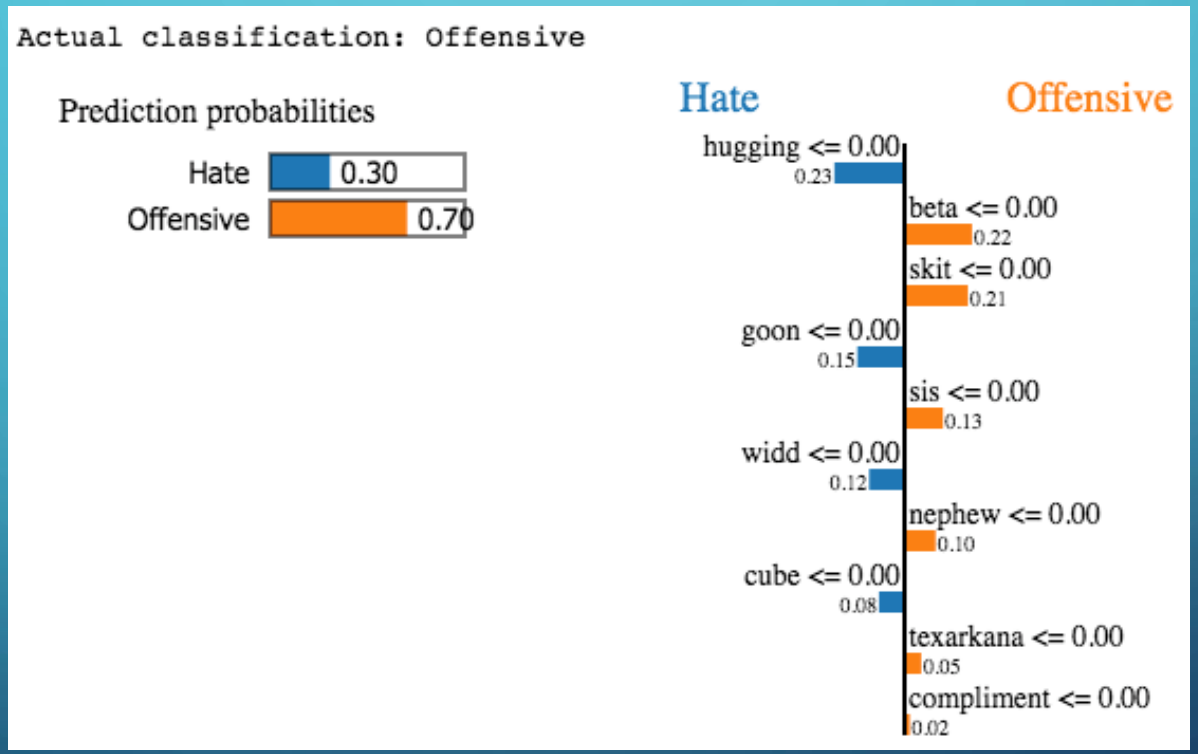
```
{'faggot': ['nigger', 'monkey', 'they', 'redneck', 'queer'],
 'nigger': ['niggers', 'monkey', 'ur', 'please', 'making'],
 'nigga': ['niggas', 'boss', 'might', 'broke', 'up'],
 'white': ['park', 'school', 'every', 'black', 'president'],
 'queer': ['his', 'lookin', 'check', 'fan', 'henny'],
 'gay': ['joe', 'fo', 'muzzie', 'looking', 'ol'],
 'coon': ['check', 'forget', 'took', 'for', 'water'],
 'kill': ['thug', 'street', 'cracker', 'internet', 'dirty'],
 'hate': ['live', 'd', 'swear', 'single', 'chinks'],
 'trash': ['people', 'man', 'faggots', 'jews', 'racist'],
 'black': ['probably', 'use', 'knows', 'filthy', 'followers'],
 'retard': ['wetbacks', 'throat', 'everyone', 'gook', 'muzzie'],
 'beaner': ['teabaggers', 'killing', 'niglet', 'inch', 'election'],
 'wetback': ['clearly', 'yelling', 'anti', 'dc', 'pops'],
 'homosexual': ['lgbtq', 'media', 'mgr', 'uwi', 'sideways'],
 'gays': ['jason', 'keri', 'texts', 'ona', 'arrested']}
```

WORD2VEC

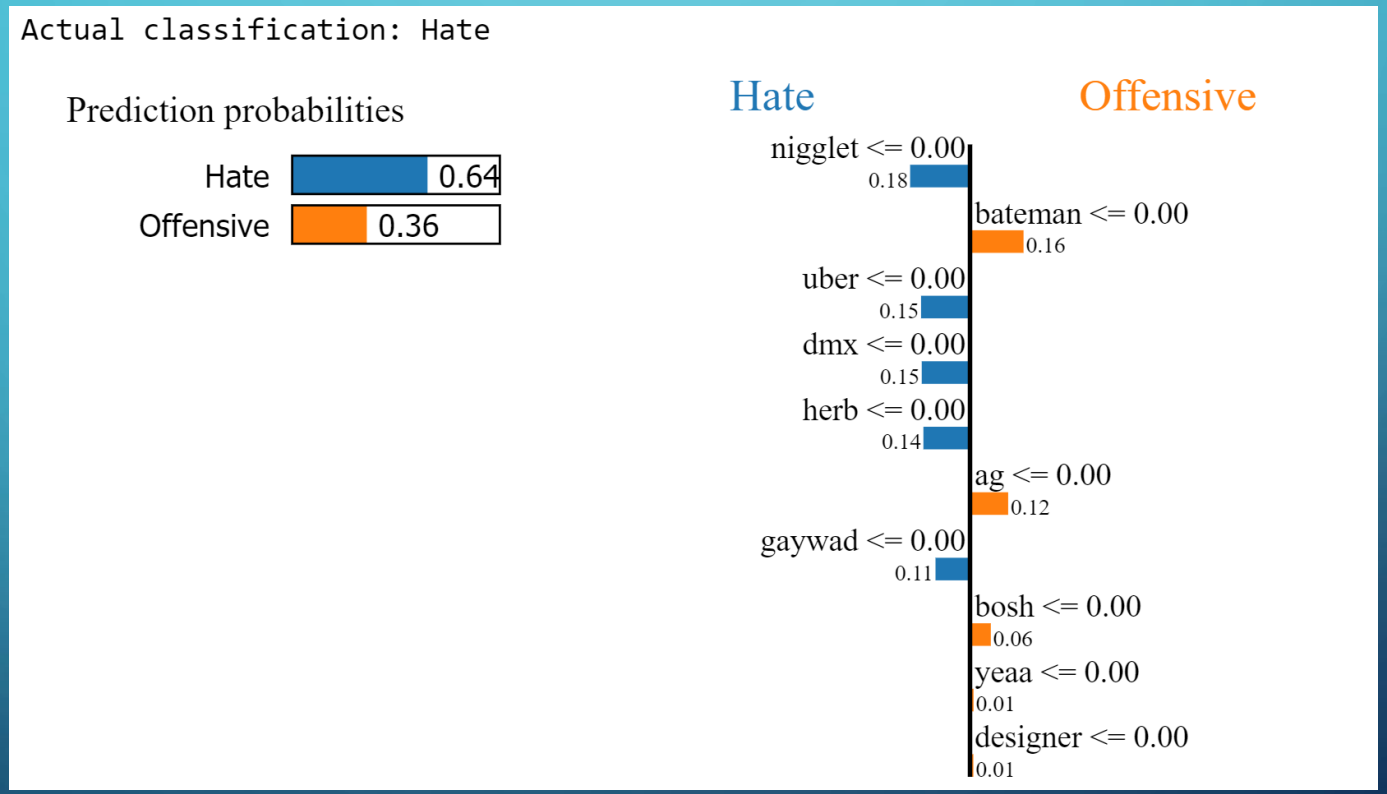


CUSTOM

LIME – SPECIFIC EXAMPLE (OFFENSIVE)



LIME – SPECIFIC EXAMPLE (HATE)



SUMMARY

- TF-IDF + extracted features appears to be best for prediction and interpretability
- Hateful words were captured by feature importance
- Differences in W2V and Custom CBOW – some semantic meaning is still captured
- Model performance already exceeds that of original authors' (80% vs 60%)
- Will use TF-IDF + extracted features moving on