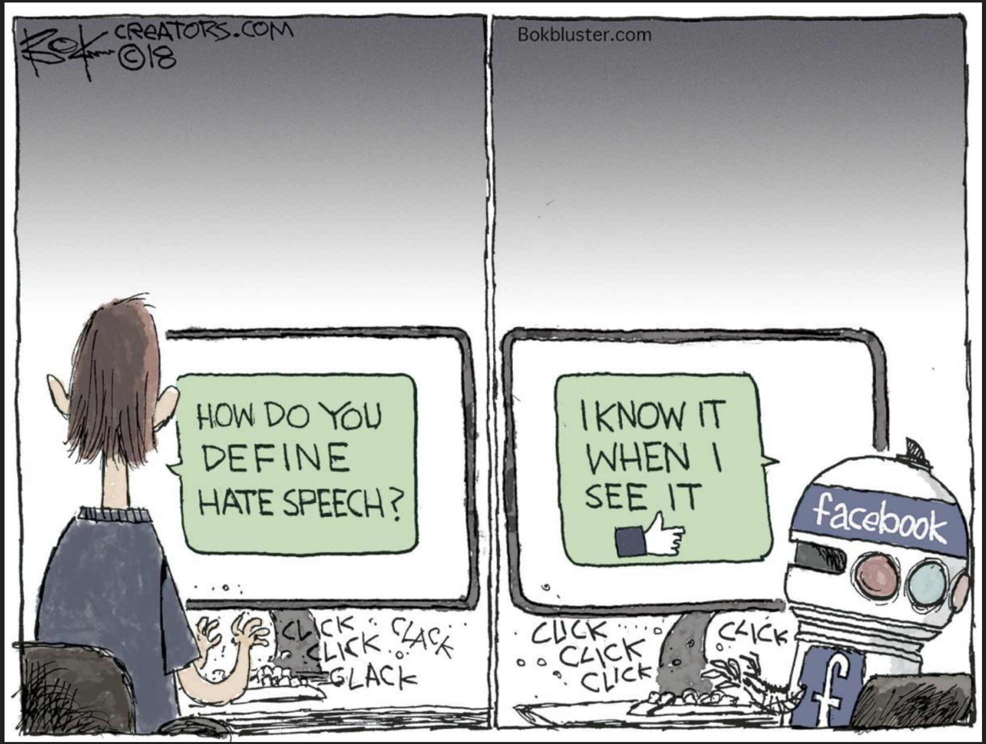
#hate speech recognition



#overview

The What

- Twitter - largest microblogging website in the world - 330 million global users
- Content is mostly unrestricted
- The Supreme Court has never created any category of speech that is defined by its hateful conduct.
- So what is hate speech?

Principal's Office
I SAID THE "S·H" WORD.
I SAID THE "F" WORD.
I SAID "CHRISTMAS."

- Hate speech is defined as abusive or threatening speech that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation

- What is considered ‘hate speech’ is a fine line between the first amendment and harmful conduct unto others

- With this project, we hope to determine what can be recognized as hate speech on a dataset of over 32,000 tweets

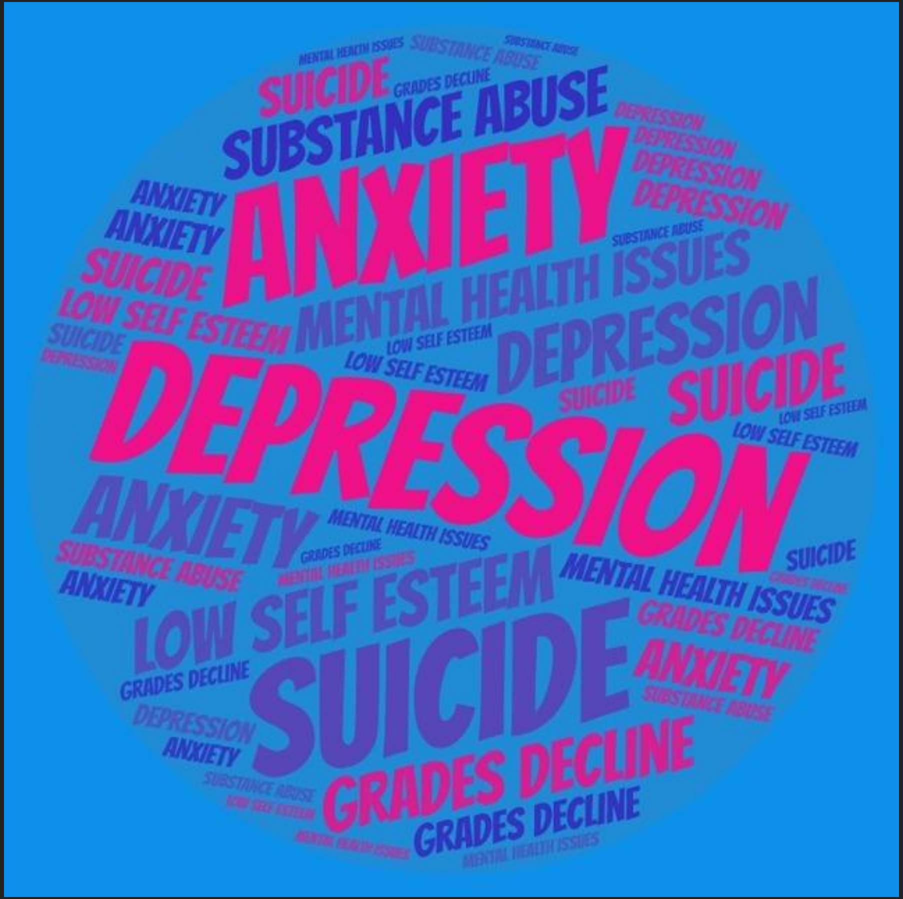
FREE SPEECH COMPLAINTS
STORY ©2010 ROCK RIVER TIMES
STORYCARTOONS.com
DEATH TO ABORTIONISTS
WHITE POWER
GAS

#inspiration
The Why

- 59% of Americans believe online hate and harassment make hate crimes more common

- 22% feel less safe in their community because of online hate

- 85% want the government to act by improving training and resources for police on cyber hate

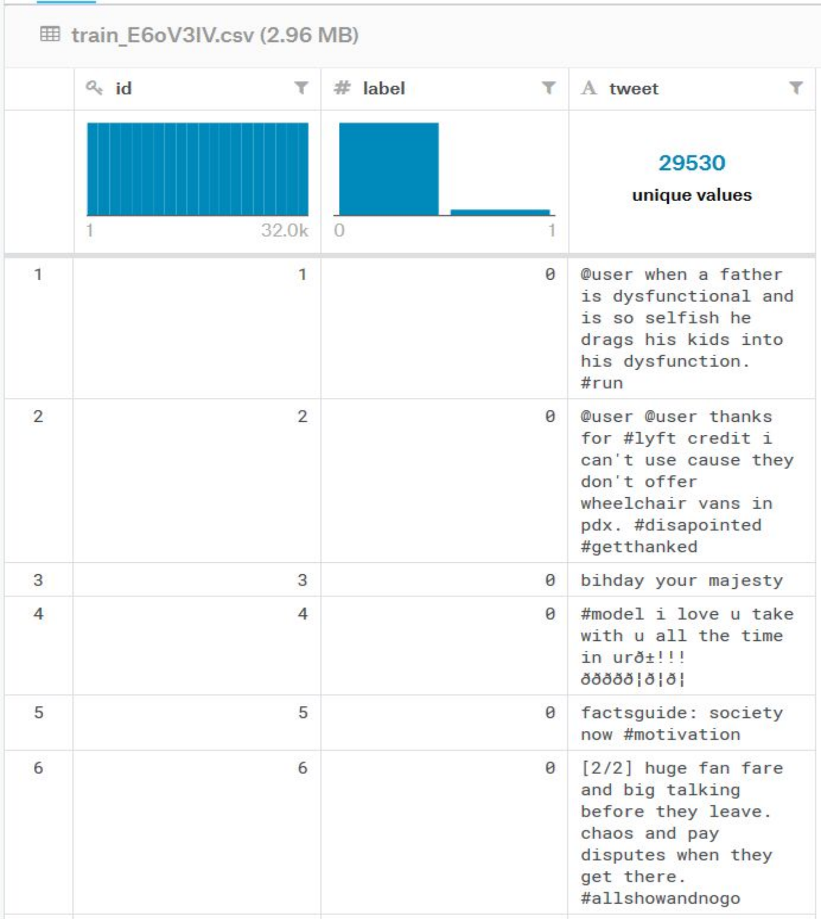


#data

- 31935 records

- 29,695 Non-Hate Labeled Twitter Data (93%)
- 2240 Hate Labeled Twitter Data (7%)

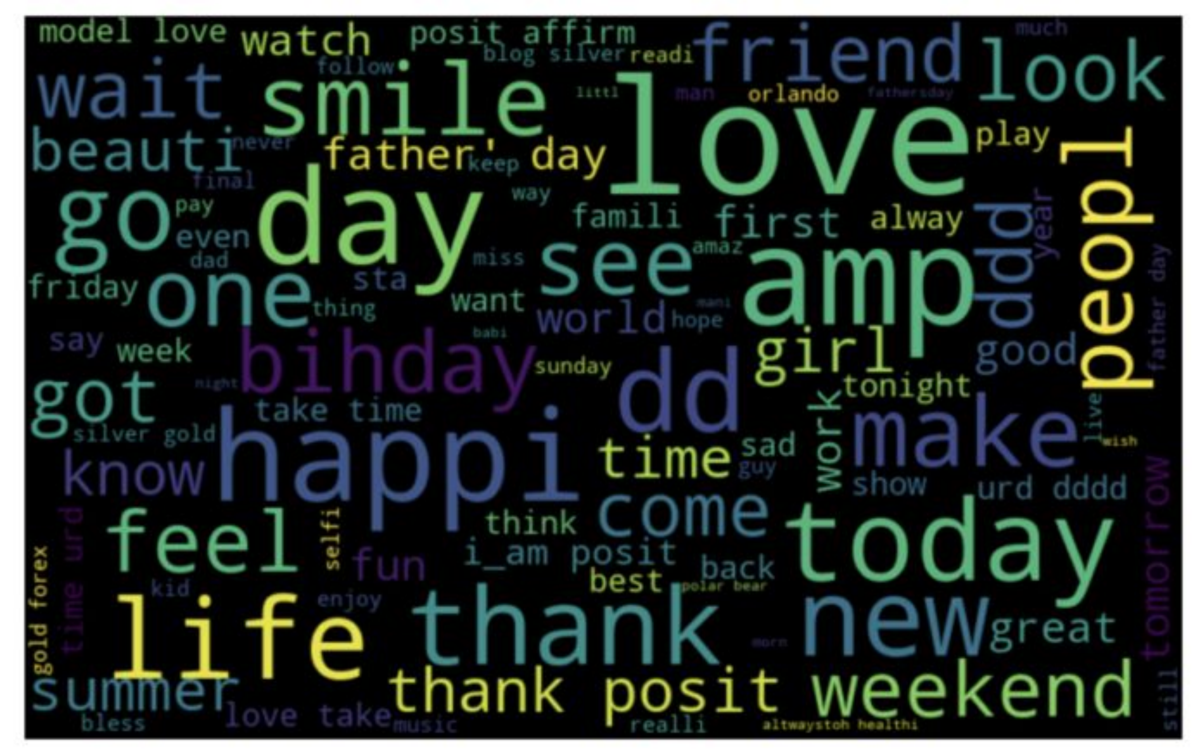
- Source: https://www.kaggle.com/vkrahul/twitter-hate-speech



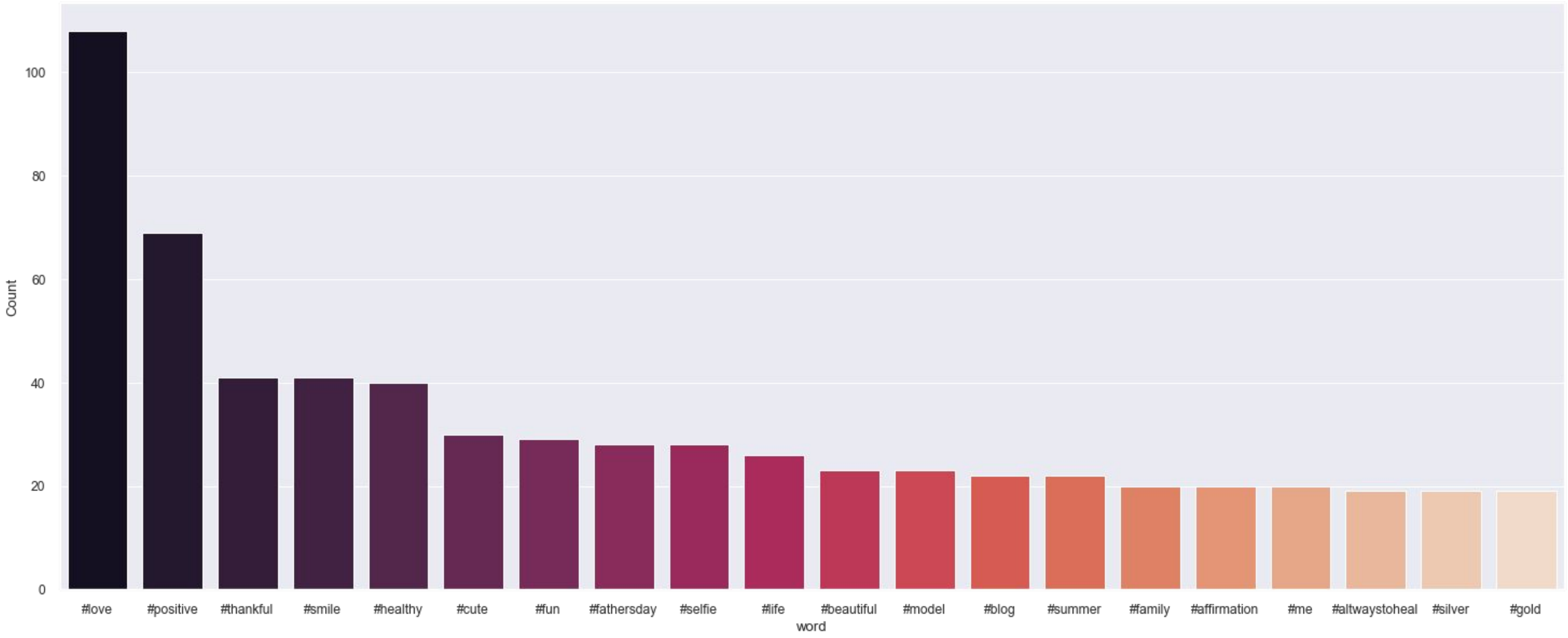
train_E6oV3IV.csv (2.96 MB)

	id	label	tweet
	1 32.0k	0 1	29530 unique values
1	1	0	@user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction. #run
2	2	0	@user @user thanks for #lyft credit i can't use cause they don't offer wheelchair vans in pdx. #disapointed #getthanked
3	3	0	bihday your majesty
4	4	0	#model i love u take with u all the time in urð±!!! ðððð¦ð¦ð¦
5	5	0	factsguide: society now #motivation
6	6	0	[2/2] huge fan fare and big talking before they leave. chaos and pay disputes when they get there. #allshowandnogo

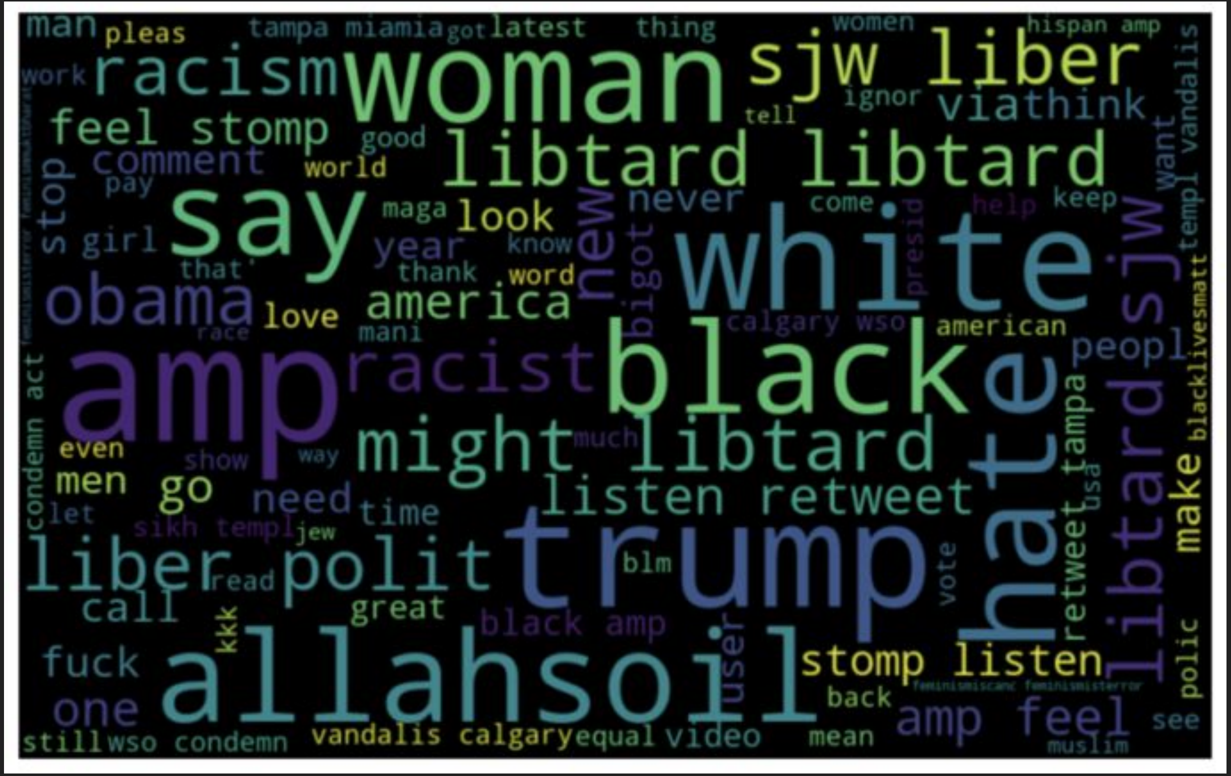
#positive_word_cloud



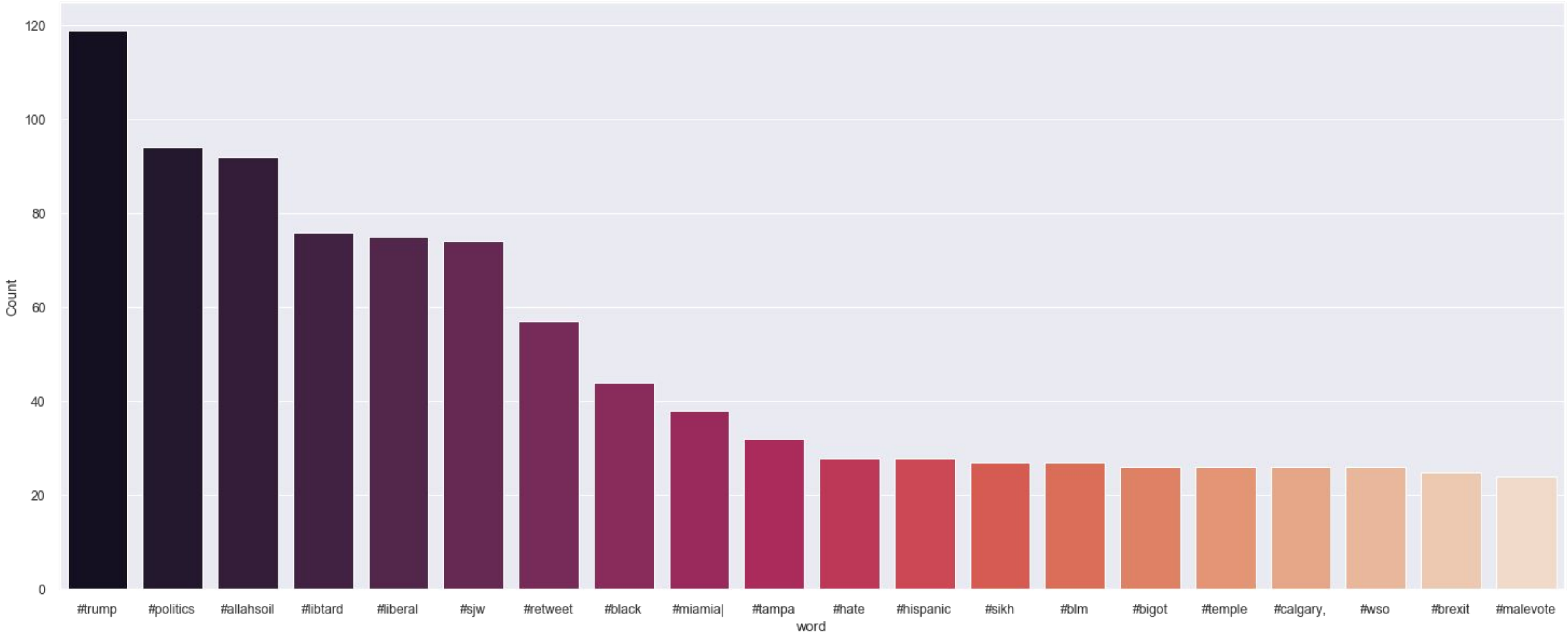
#top20_positive_hashtags



#negative_word_cloud



#top20_negative_hashtags

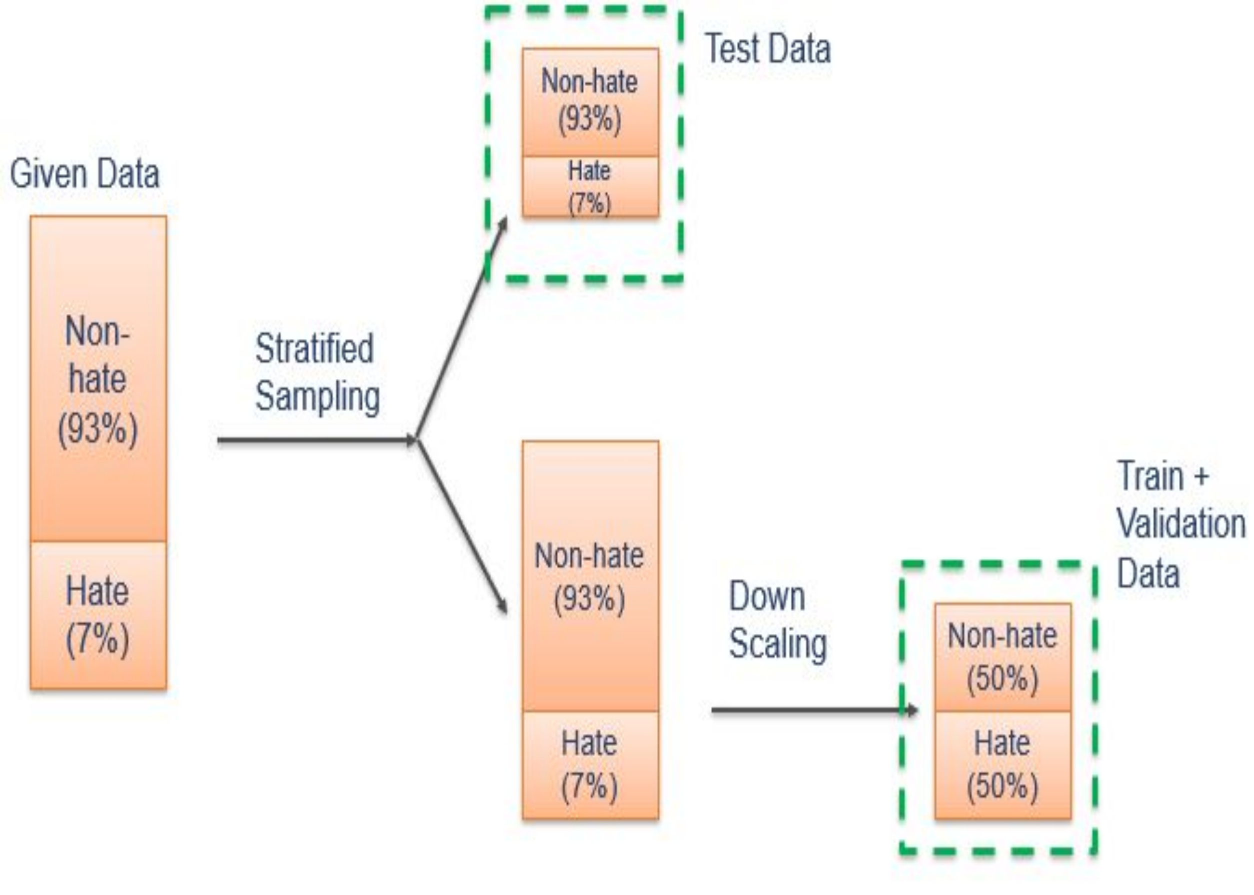


The #how

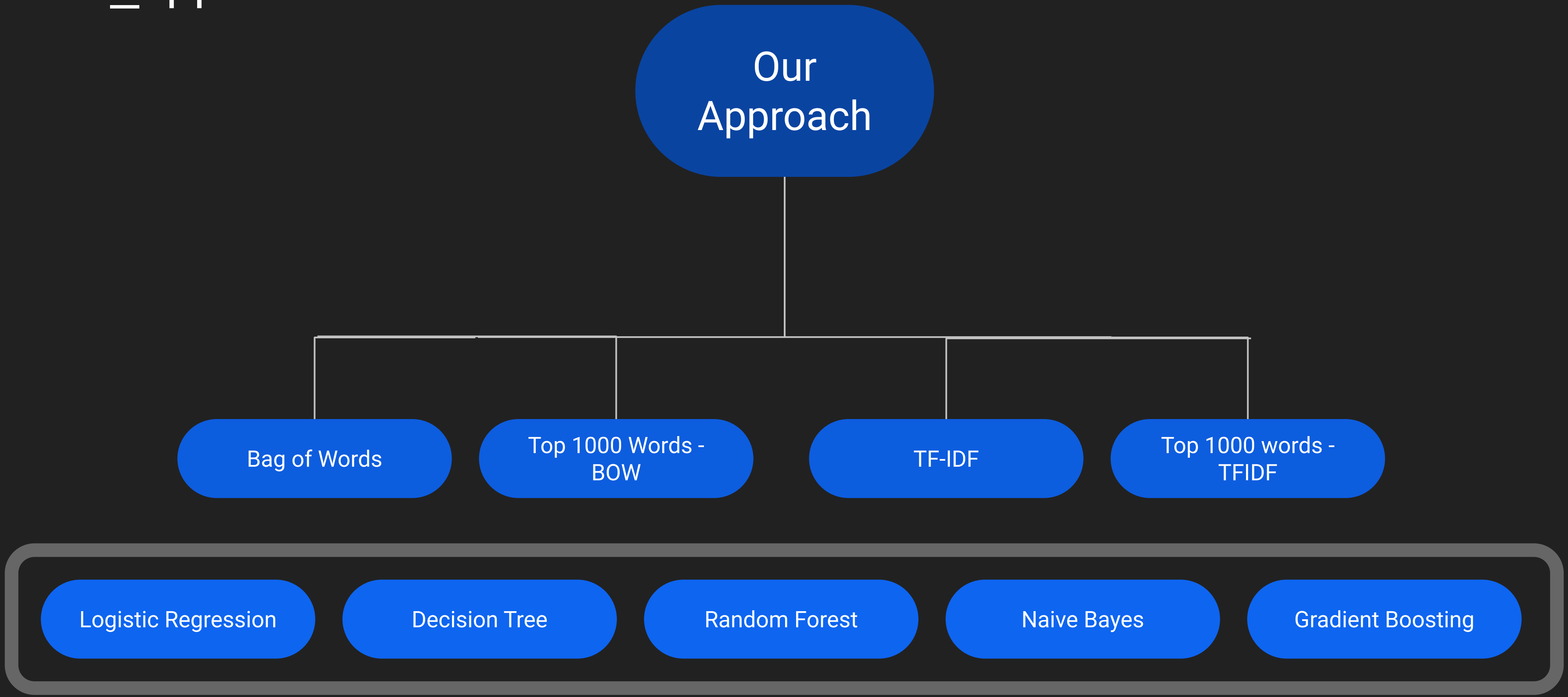
#data architecture

- We had 93% of Non-Hate tweets and 7% Hate tweets

- We downscaled to remove the bias



#our_approach



Our Approach

Bag of Words

Top 1000 Words - BOW

TF-IDF

Top 1000 words - TFIDF

Logistic Regression

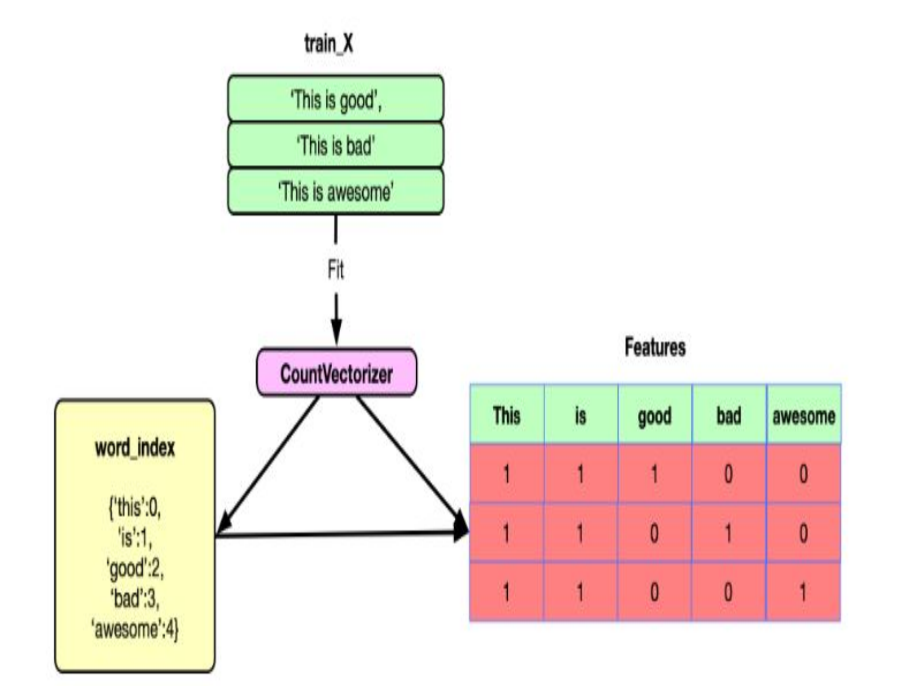
Decision Tree

Random Forest

Naive Bayes

Gradient Boosting

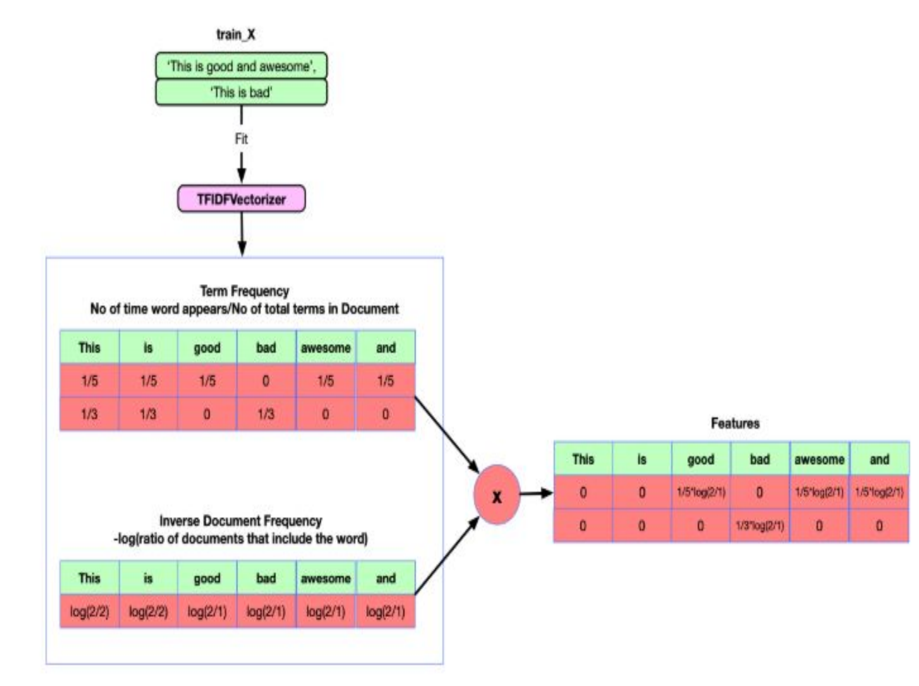
#bag_of_words



Features

This	is	good	bad	awesome
1	1	1	0	0
1	1	0	1	0
1	1	0	0	1

#term_frequency_inverse_document _frequency



Term Frequency
No of time word appears/No of total terms in Document

This	is	good	bad	awesome	and
1/5	1/5	1/5	0	1/5	1/5
1/3	1/3	0	1/3	0	0

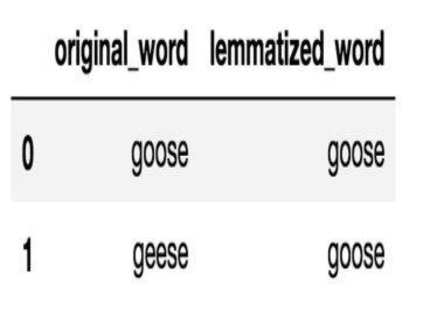
Inverse Document Frequency
-log(ratio of documents that include the word)

This	is	good	bad	awesome	and
log(2/2)	log(2/2)	log(2/1)	log(2/1)	log(2/1)	log(2/1)

Features

This	is	good	bad	awesome	and
0	0	1/5*log(2/1)	0	1/5*log(2/1)	1/5*log(2/1)
0	0	0	1/3*log(2/1)	0	0

#data #cleaning

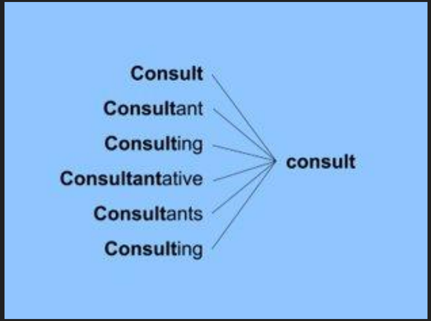


	original_word	lemmatized_word
0	goose	goose
1	geese	goose

#LEMMATIZATION

ASAP
FOMO
BTW
YOLO
THX
DOE
DEAD!
REKT
LOL
BAE
WTF
Yaaaas!!
TURNT
OMG
vexels

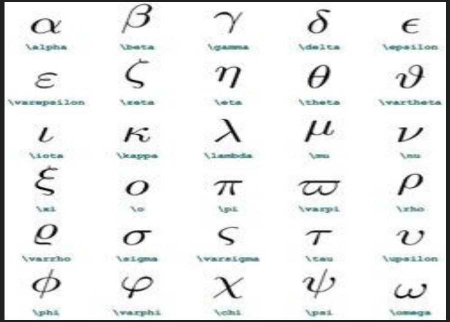
#SLANG_WORDS



Consult
Consultant
Consulting
Consultantative
Consultants
Consulting
consult

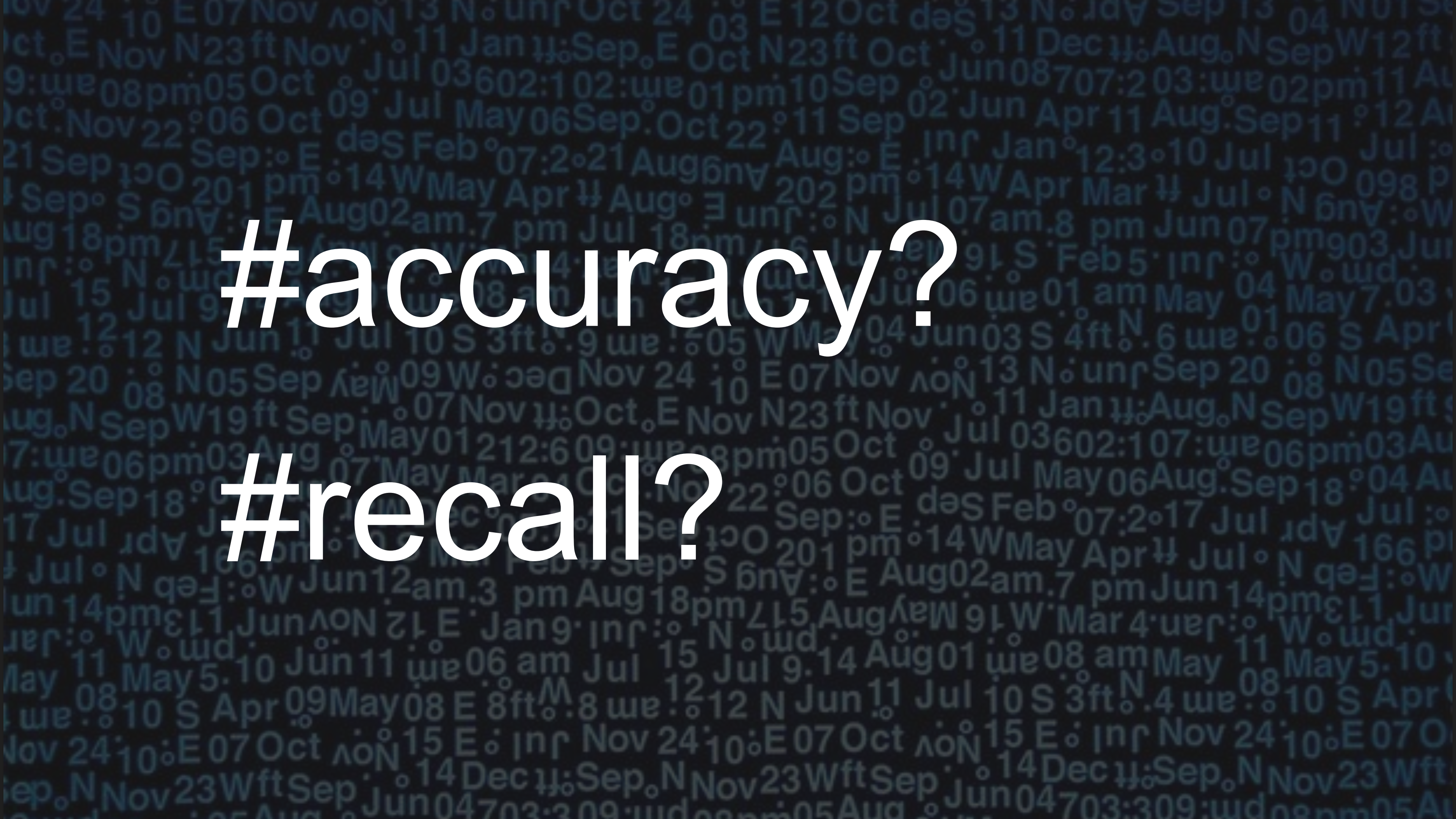
#STEMMING

#STOPWORDS



α \alpha	β \beta	γ \gamma	δ \delta	ϵ \epsilon
ε \varepsilon	ζ \zeta	η \eta	θ \theta	ϑ \vartheta
ι \iota	κ \kappa	λ \lambda	μ \mu	ν \nu
ξ \xi	o \o	π \pi	ϖ \varpi	ρ \rho
ϱ \varrho	σ \sigma	ς \varsigma	τ \tau	υ \upsilon
ϕ \phi	φ \varphi	χ \chi	ψ \psi	ω \omega

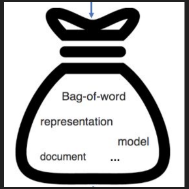
#GREEK_CHARACTERS

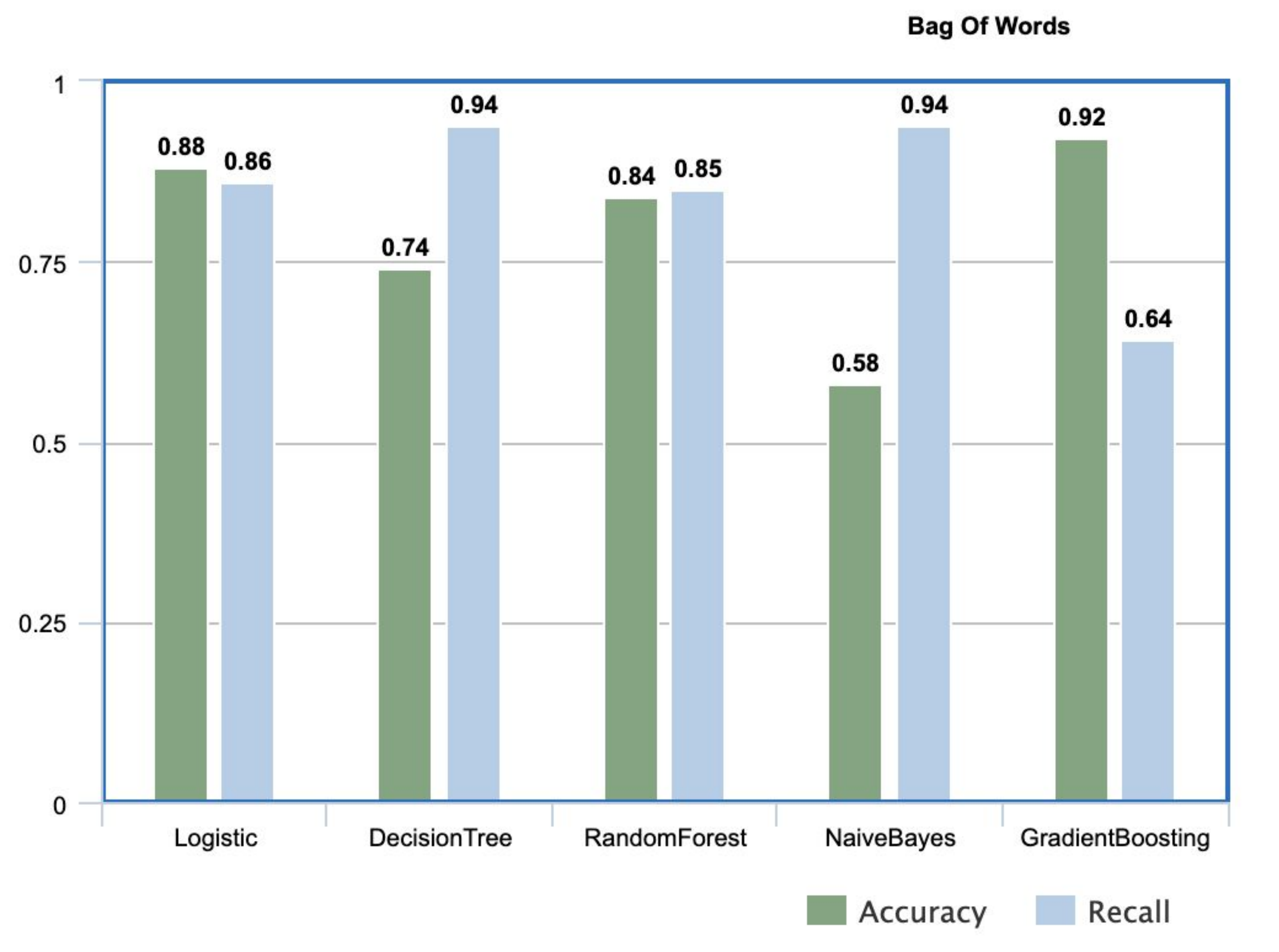


#accuracy?

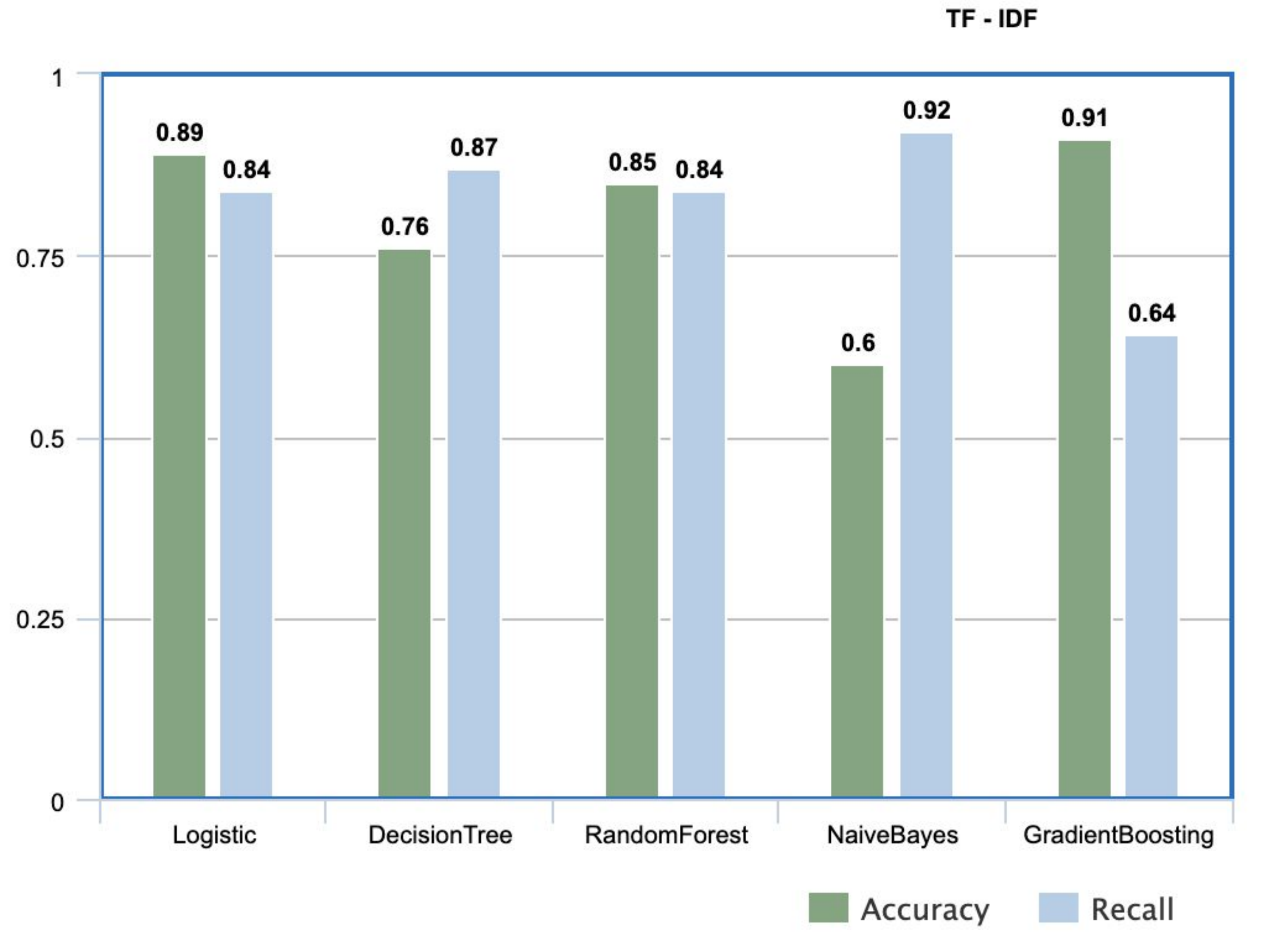
#recall?

#Bag_of_Words


Bag-of-word
representation
model
document
...

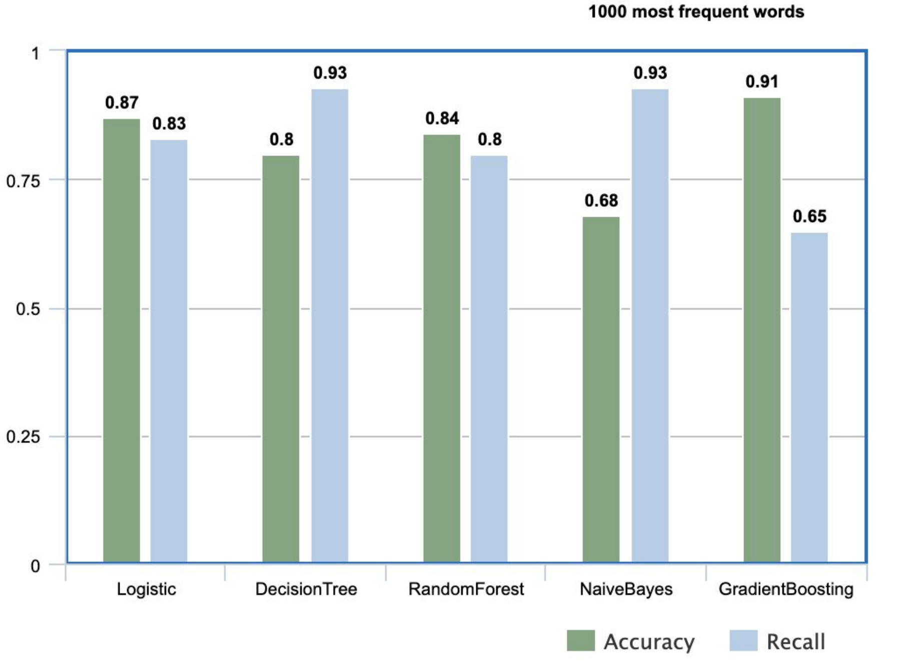

Bag Of Words
1
0.75
0.5
0.25
0
0.88
0.86
0.74
0.94
0.84
0.85
0.58
0.94
0.92
0.64
Logistic
DecisionTree
RandomForest
NaiveBayes
GradientBoosting
Accuracy
Recall

#TF_IDF

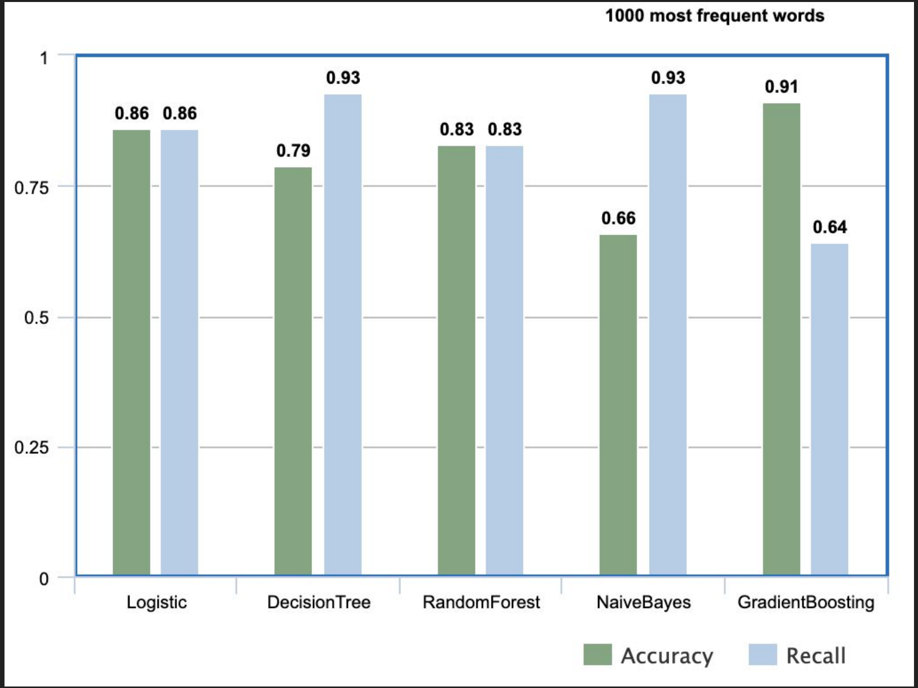

TF - IDF
1
0.75
0.5
0.25
0
0.89
0.84
0.76
0.87
0.85
0.84
0.6
0.92
0.91
0.64
Logistic
DecisionTree
RandomForest
NaiveBayes
GradientBoosting
Accuracy
Recall

#1000_most_frequent_words

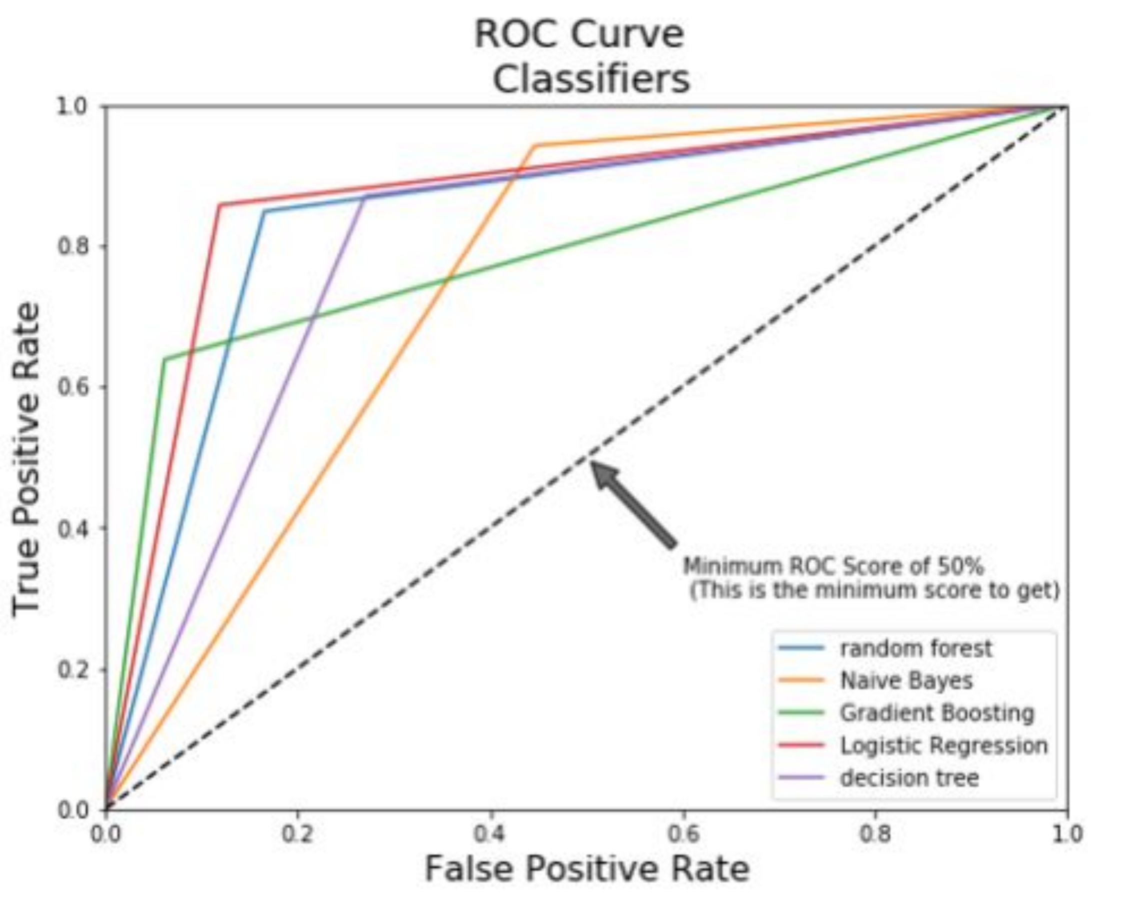
#TF_IDF



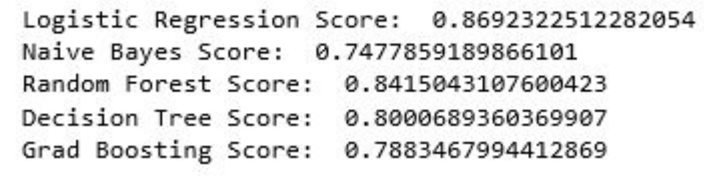
#Bag_Of_Words



#ROC_curves_for_Bag_of_Words_approach



AUC



```
Logistic Regression Score:  0.8692322512282054
Naive Bayes Score:  0.7477859189866101
Random Forest Score:  0.8415043107600423
Decision Tree Score:  0.8000689360369907
Grad Boosting Score:  0.7883467994412869
```

#key_insights

- Politics, race and sexual tweets form a major chunk of hate tweets
- Results obtained when we played around with an imbalanced data set were inaccurate. The model predicts new data to belong to the major class in the training set due to the skewed nature of the train data.
- The weighted accuracy of a classifier is not the only metric to be looked at while evaluating the performance of the model. Business context plays a vital role as well.

#challenges #future_steps

- Change in attitude towards topics over time and historical context.

  "The Nazi organization was great."

- Automatic hate-speech detection is a closed-loop system:

  Hate verbiage is not restricted to text. Future prospects include recognising hate portrayed through pictures.

#team

#adwait

#vedant

#sushrut

#sneha

#rishab

#tanvir

#chavi