HATE SPEECH DETECTION

ITAY ALUSH
GAL EICHENBAUM
BEN TOBIN

NLP & TEXT MINING FINAL PROJECT | DR. LEV MUCHNIK | 19.07.2023

RACISM

ANTISEMITISM

GENDER PREJUDICE

ISRAEL CONFLICT

Montana | February 17, 2023 | Sexual Orientation

— ▸ Montana Man Convicted for Attacks on Local LGBTQ Community

California | February 17, 2023 | Religion

— ▸ California Man Charged for Allegedly Shooting Two Jewish Men in Los Angeles

New Jersey | February 1, 2023 | Religion

— ▸ Passaic County Man Arrested for Attempt to Firebomb Synagogue

South Carolina | February 1, 2023 | Gender Identity

— ▸ Two Men Charged with Hate Crimes and Obstruction in the Murder of Transgender Woman

Idaho, Oregon, Washington | January 30, 2023 | Race, Ethnicity

— ▸ Four Men Sentenced for Hate Crime and False Statement Charges After Racially-Motivated Assault

Louisiana | January 25, 2023 | Sexual Orientation

— ▸ Louisiana Man Sentenced to 45 Years for Kidnapping and Attempting to Murder a Gay Man

Florida | January 25, 2023 | Race

— ▸ Two Florida Men Sentenced for Hate Crime Following Racially-Motivated Assault

Idaho| January 12, 2023 | Sexual Orientation

— ▸ Idaho Man Indicted for Federal Hate Crime Against LGBTQ Residents of Boise

Washington | December 14, 2022 | Religion

— ▸ Washington Man Indicted for Arsons at Jehovah's Witness Kingdom Halls

Missouri | December 13, 2022 | Religion

— ▸ Missouri Man Pleads Guilty to Burning Down Islamic Center

Austria

An increased number of hate crimes took place this year. The NGO SoHo collected those between January — July 2021. For instance, a group of young people were assaulted in Vorarlberg, and the victims were hospitalised with serious injuries. Rainbow flags and other symbols were vandalised.

Cyprus

Pressured by homophobic bullies at a party to binge drink beyond his tolerance, a teen fainted and was left to choke to death in February. Three of the peers received 18 months of suspended sentence for involuntary manslaughter.

Denmark

"Live and Let Live" published 1,000 accounts of antiLGBTQI hate crimes and speech.

France

A gay man was murdered in April, a lesbian couple in August, and a trans migrant sex worker woman in September. At least six trans people are known to have committed suicide due to transphobic harassment.

55% of hate crimes in Sweden have racial motivations: Report

ADL Tracker
@ADL_Tracker

A fan of the British soccer team, Arsenal, has pleased guilty to shouting "Hitler should have finished the job" at an Arsenal match against rival, Tottenham. As a result, he has been banned from soccer games for three years and must pay 471 pounds in fines.... https://t.co/lXknKmx9Jy

July 12, 2023

Cameroon [edit]

In February 2019, deputy justice minister Jean de Dieu Momo advanced an antisemitic canard during prime time on Cameroon Radio Television, and suggested that Jewish people had brought the holocaust upon themselves.[3][4]

Overview of incidents reported by other sources

Violent attacks against people	Threats	Attacks against property	Total
997	246	741	1984

Table 1: 2013 official figures on reported racist crimes and complaints[15]

Austria	110
Bulgaria	Not available
Croatia	33
Cyprus	8
Czech Republic	186
Denmark	Not available
Estonia	Not available
Finland	833
France	1,376
Germany	5,131
Greece	43
Hungary	3
Iceland	0
Italy	194
Ireland	93
Latvia	22
Lithuania	84
Luxembourg	31
Malta	Not available
Netherlands	Not available
Poland	719
Romania	Not available
Slovakia	Not available
Spain	384
Sweden	1,733
England & Wales	30,788
Scotland	4,735
Northern Ireland	704

Source: OSCE/ODIHR and ENAR questionnaire responses

2019 – White supremacist shootings in El Paso, Texas

2018 – Pittsburgh synagogue shooting is deadliest attack on Jews in U.S. history

2018 – Dehumanization of Immigrant Children at the Border

2016 – Pulse nightclub shooting by domestic terrorist inspired by ISIS

2015 – Charleston Church Massacre

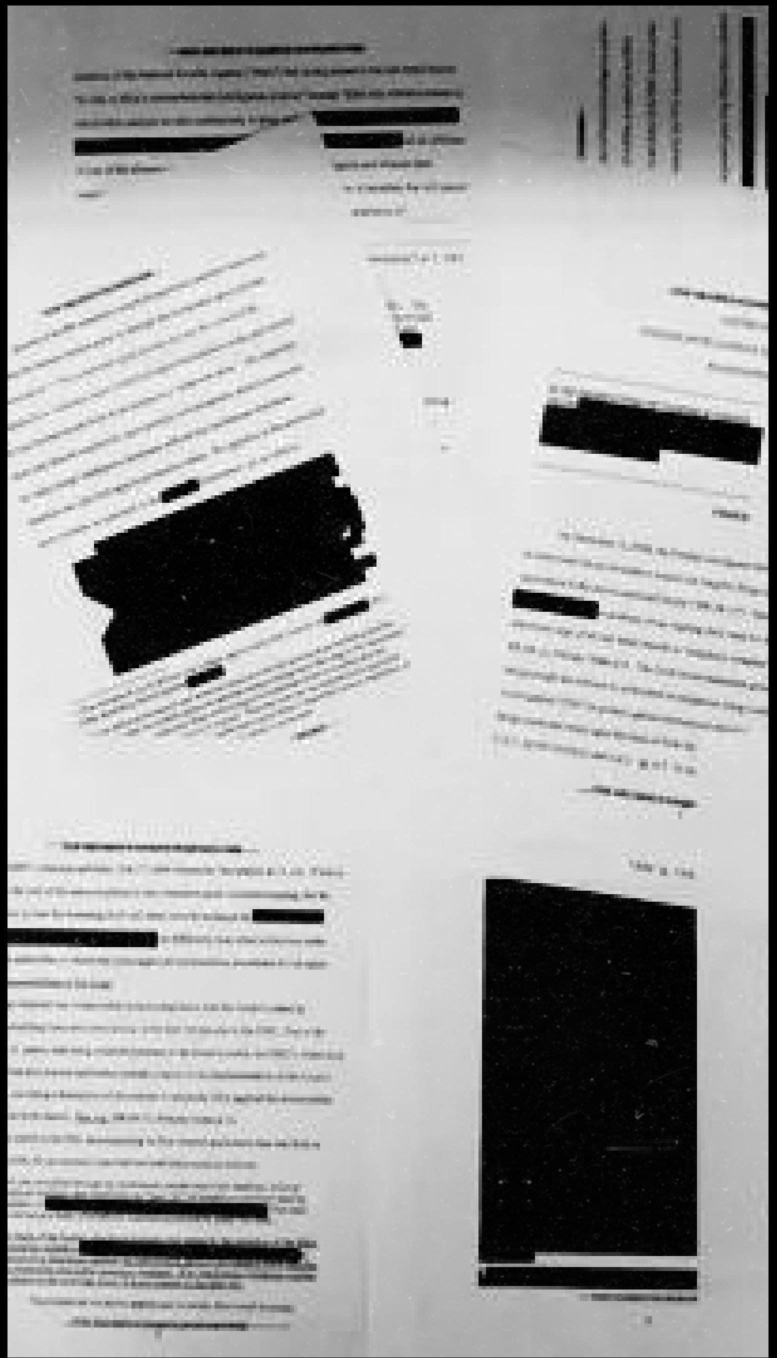
2012 – Sikh Gurdwara shooting in Wisconsin

RESEARCH QUESTION

Can we Identify hate speech using M.L models?

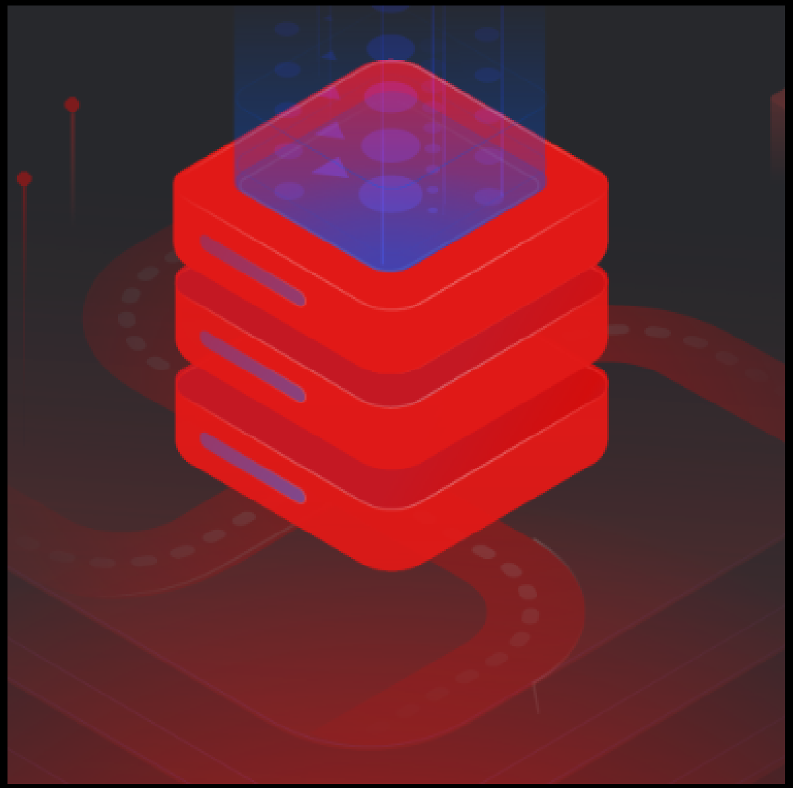
BUSINESS APPLICATION

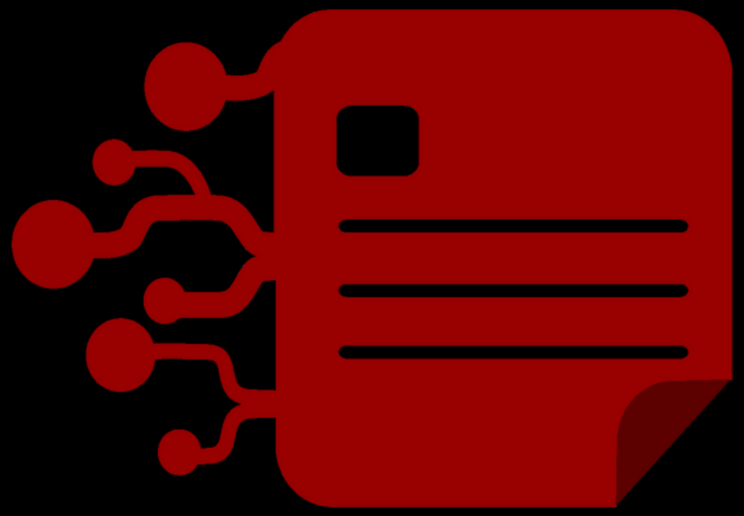
CAN WE HELP SOCIAL MEDIA COMPANIES REDUCE THE AMOUNT OF HATE SPEECH ON THEIR PLATFORMS?

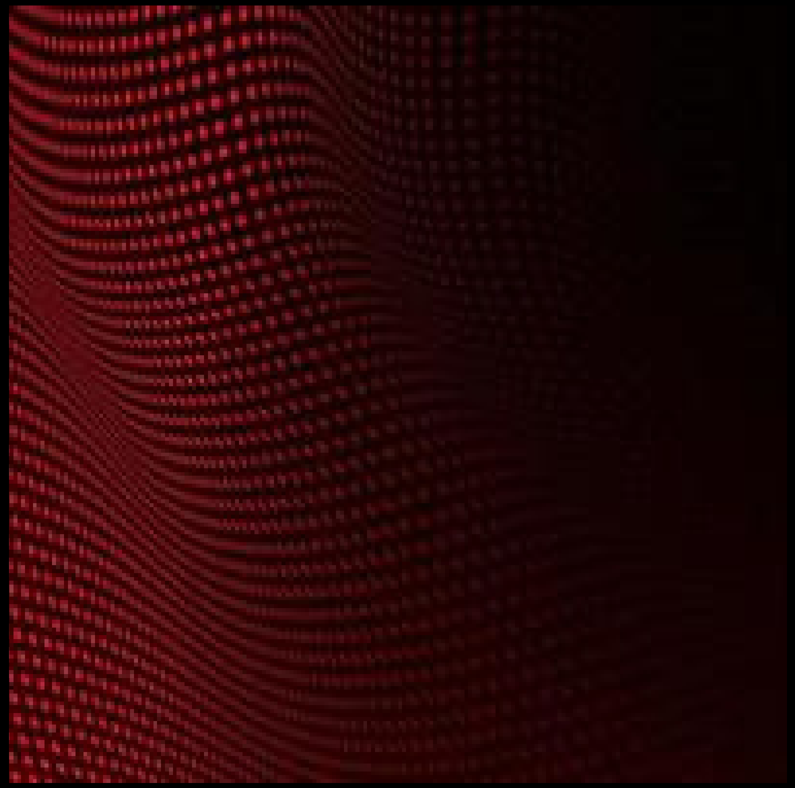


OUR PROJECT STAGES

- DATA + EDA
- PREPROCCESSING
- NLP ANALYSIS
- CLASSIFICATION







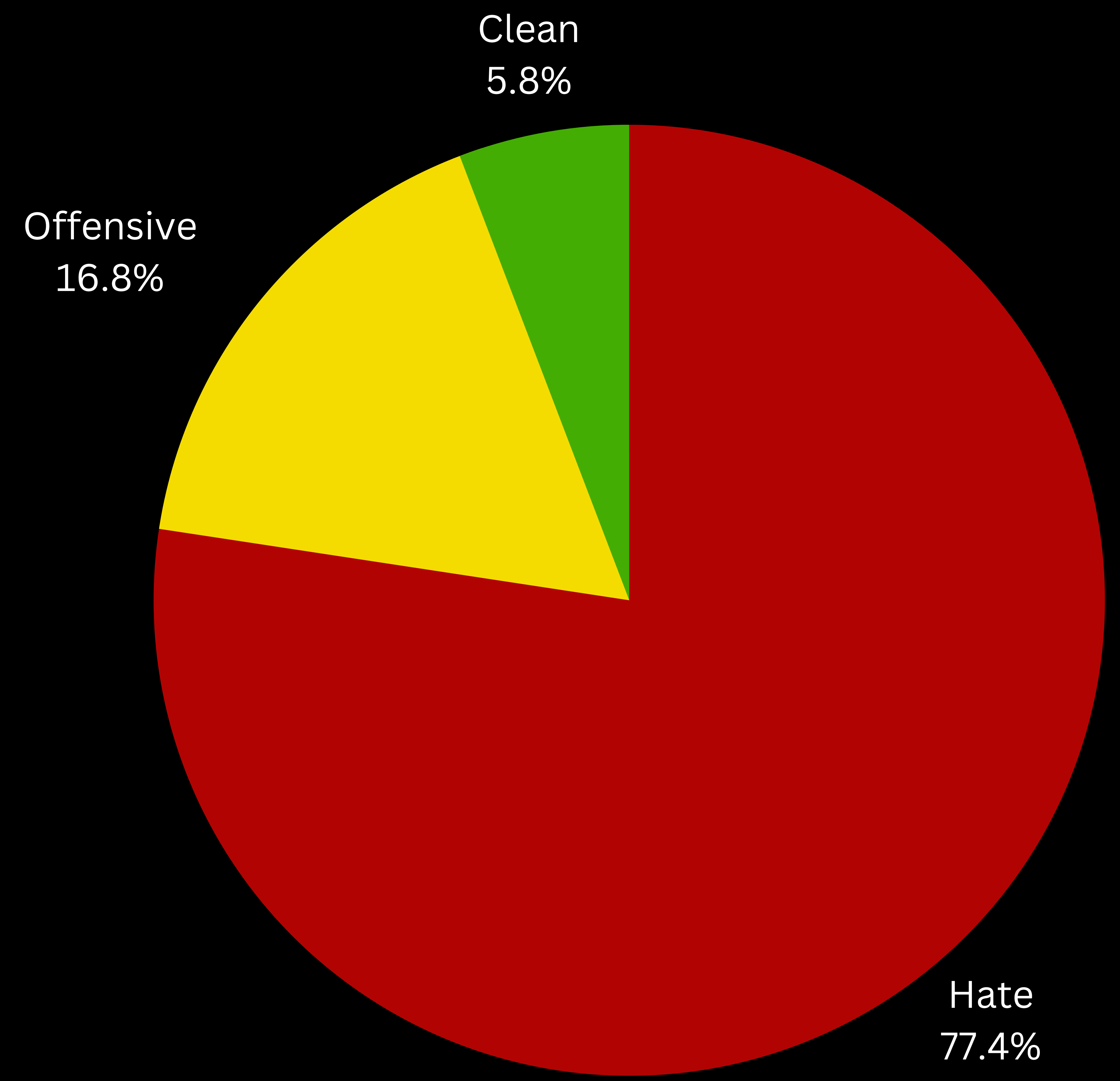
DATA

- Tweets collected by a group of researches and saved in a csv file.
- The tweets were shown to various people and asked to label 1 of three categories.
- Category with most votes decided final label.
- Total of 24,783 tweets

HATE

OFFENSIVE

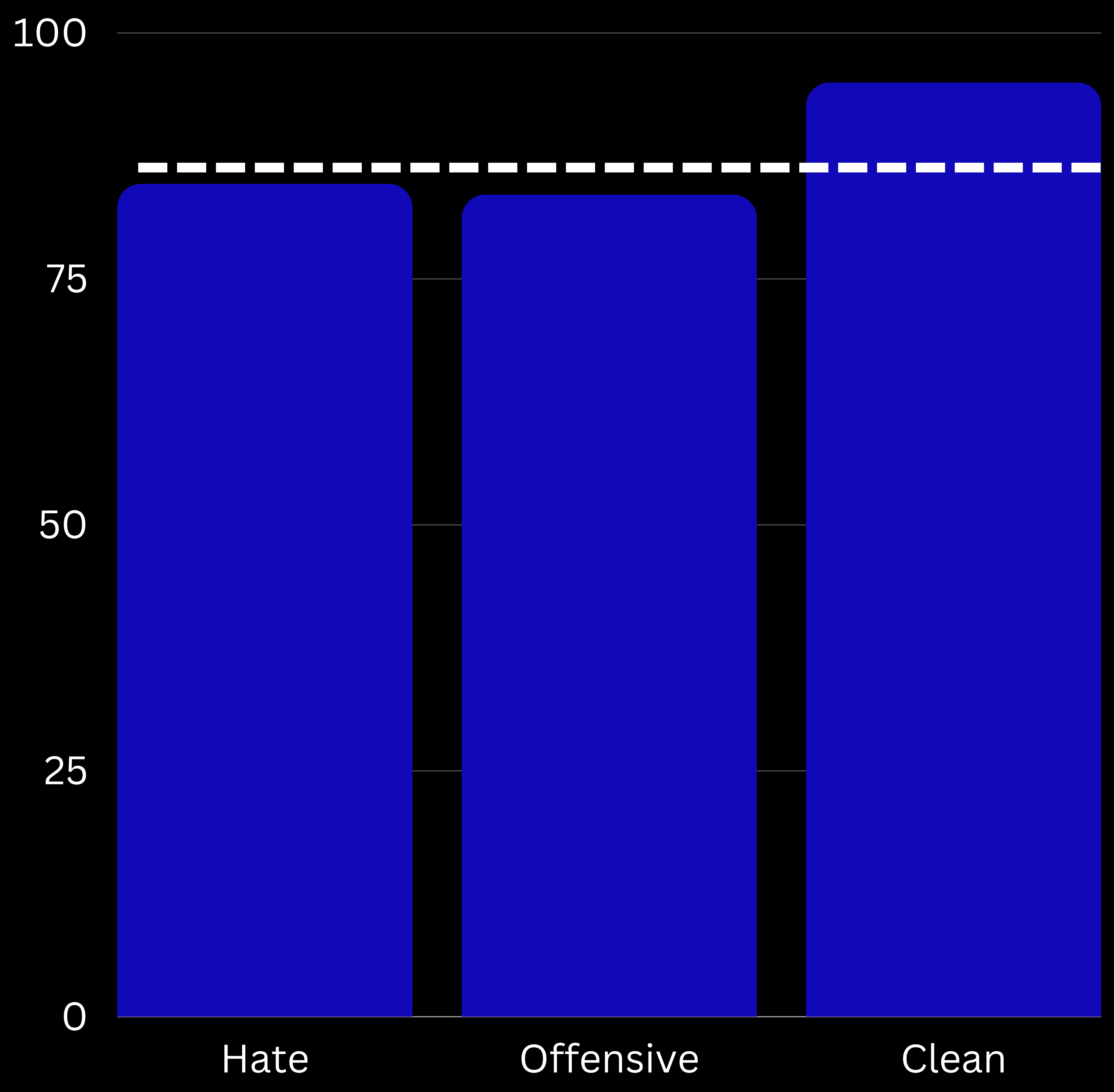
CLEAN



CLASS DISTRIBUTION

Highely skewed data therfore baseline model is at ~78% accuracy

AVERAGE TWEET LENGTH PER CLASS



WORDCLOUD BY LABEL

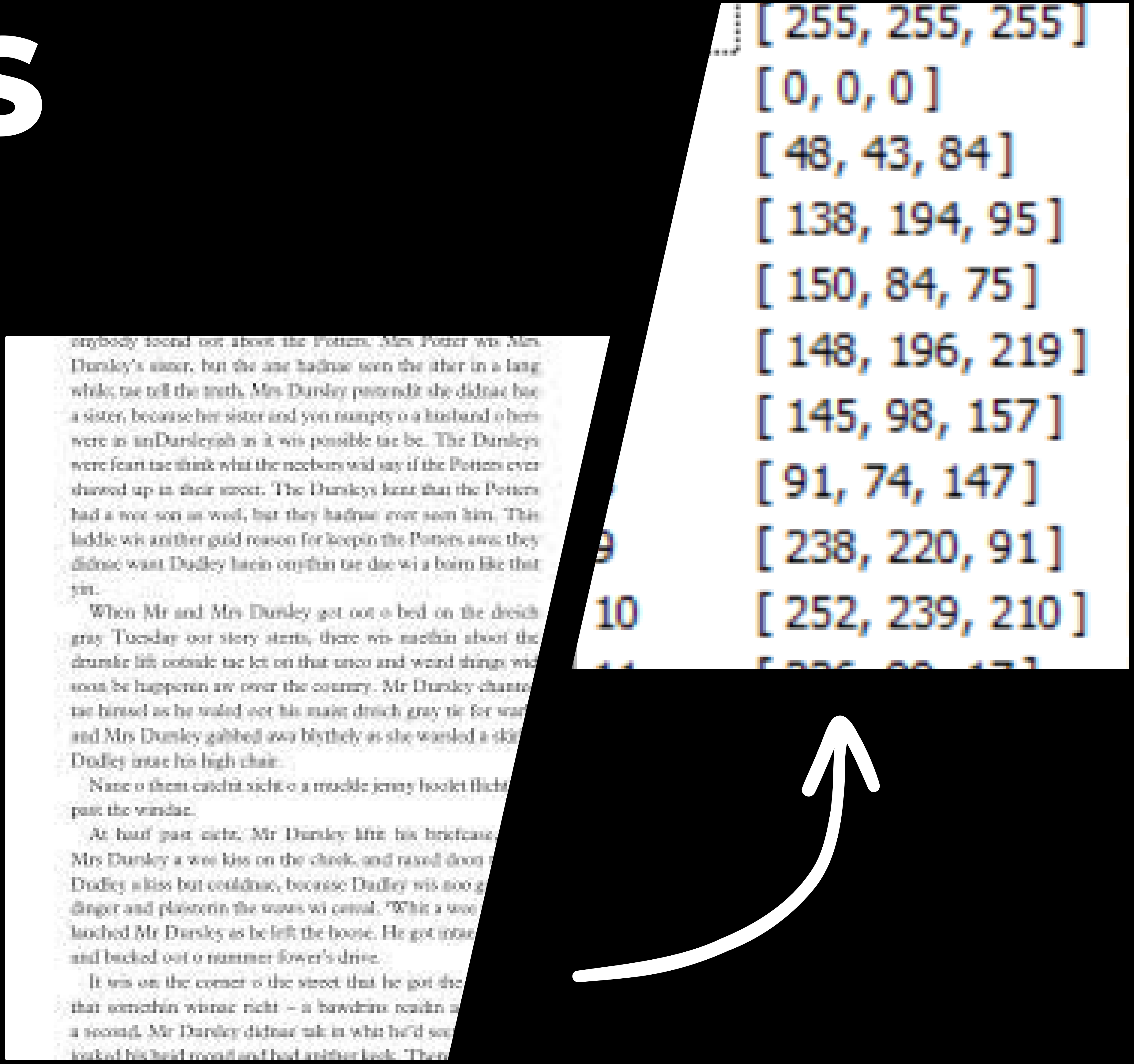
Hate Speech

Offensive

Clean

NLP ANLYSIS

- TFIDF
- LDA
- WORD2VEC
- DOC2VEC



TFIDF

- Preprocessed the tweets corpus.
- Present each tweet with a 'feature vector' and create TF-IDF matrix using Sklearns TDIDF vectorizer.
- Converted the matrix into a dataframe.
- Added the tweets labels.

During classification ran into 'Memory Error' due to large input size.

- Researched methods to lower memory allocation.
- Decided on changing Max_Feature Paramter

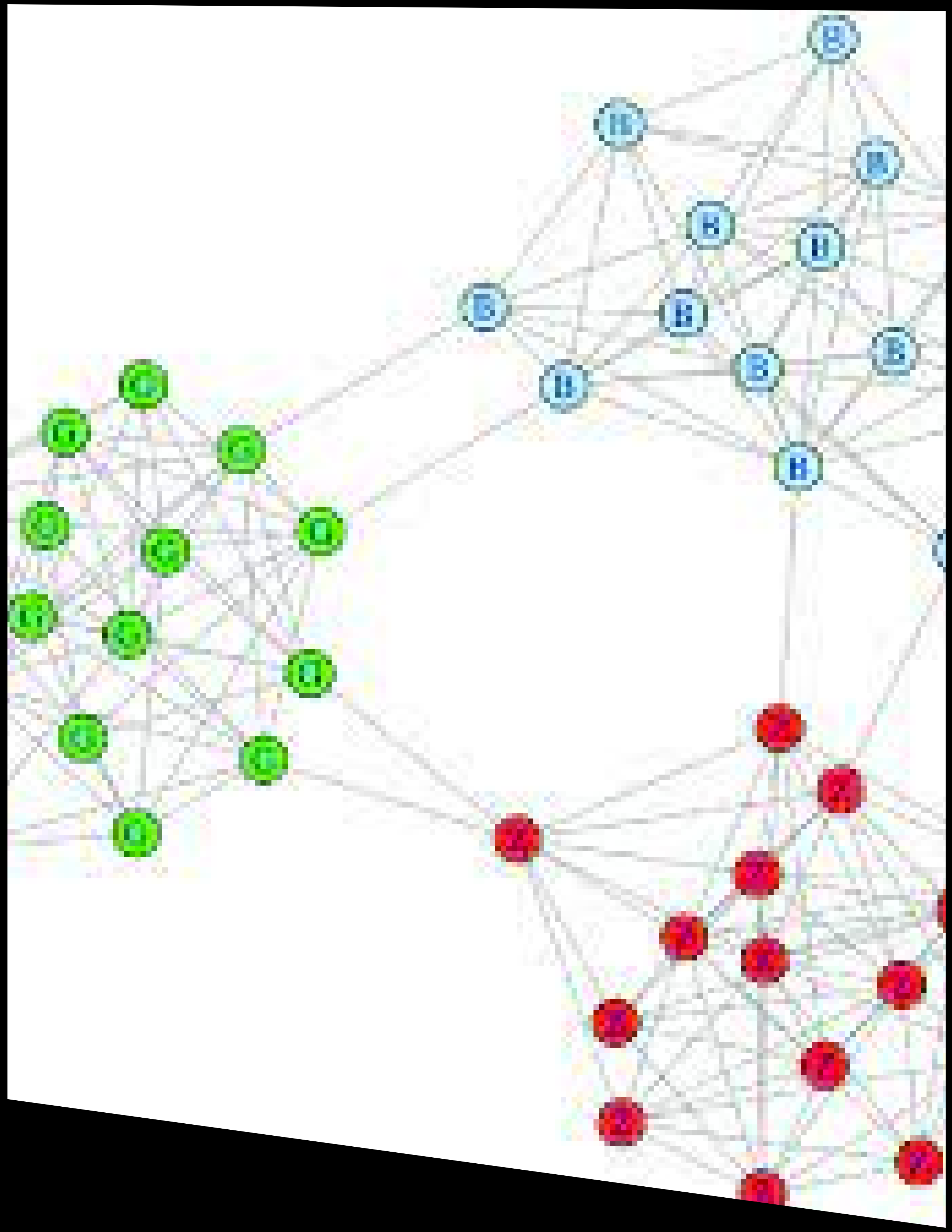
FEATURES

TWEETS

Tweet	Word 1	....	Word 1000
1			
.....			
24,783			

LDA

- LDA (Latent Dirichlet Allocation) is a powerful technique for topic modeling and text classification.
- Widely used in natural language processing tasks, including tweet classification.
- By uncovering latent topics in tweets, the LDA model predicts the topic or class of a tweet.



LDA - GRIDSEARCH

```python
# Iterate over hyperparameter combinations
for min_cf in min_cf_values:
    for rm_top in rm_top_values:
        for latent_topics in latent_topics_values:
            for topics_per_label in topics_per_label_values:
                # Create a PLDA model with current hyperparameters
                lda_model = tp.PLDAModel(
                    min_cf=min_cf,
                    rm_top=rm_top,
                    latent_topics=latent_topics,
                    topics_per_label=topics_per_label,
                    seed=42
                )

                # Add documents to the model
                for document in train_corpus:
                    lda_model.add_doc(document)

                # Train the model
                lda_model.train(100)

                # Compute perplexity
                perplexity = lda_model.perplexity

                # Check if current model has lower perplexity
                if perplexity < best_perplexity:
                    best_perplexity = perplexity
                    best_lda_model = lda_model
```

Here we tried to optimize our LDA model by its arguments.
For each argument we took a scale of possible values and trained a new model on each different combination of arguments.

At the end we saved the best LDA model with its optimized arguments.

BEST LDA MODEL PERFORMANCE

Final LDA model accuracy

LDA Model Accuracy 0.74

−/+

Top 10 words for each class

	Topic 0_word	Topic 0_value	Topic 1_word	Topic 1_value	Topic 2_word	Topic 2_value
1	the	0.069760	a	0.052698	bitch	0.043976
2	rt	0.028435	you	0.033810	i	0.033137
3	of	0.026961	i	0.032209	rt	0.026512
4	in	0.026834	bitch	0.024146	you	0.021988
5	a	0.023570	rt	0.022090	a	0.020487
6	and	0.021484	to	0.019050	bitches	0.018386
7	is	0.021379	is	0.014761	my	0.018120
8	to	0.019820	the	0.013483	hoes	0.017352
9	for	0.013691	and	0.013277	to	0.016130
10	trash	0.013101	that	0.011882	me	0.014218

TOP

WORD2VEC

(1)
Initialized and trained various parameter combos

(2)
Compared the models using the wordsim535 file

{'vector_size': , 'window': , 'min_count': , 'epochs': , 'sg': , 'min_alpha': }

MODEL	SPEARMAN	PEARSON
Default	0.01438	-0.02325
Large Vector	0.02458	-0.06847
Small Vector	0.0534	0.00854
Wide Window	0.04792	-0.02814

MODEL	SPEARMAN	PEARSON
High Min_Count	0.1	0.1984
Many Epochs	0.196	0.223
Skip Gram	0.0378	0.0232
High LR	0.0992	0.1285

WORD2VEC

(3)

Checked that model works

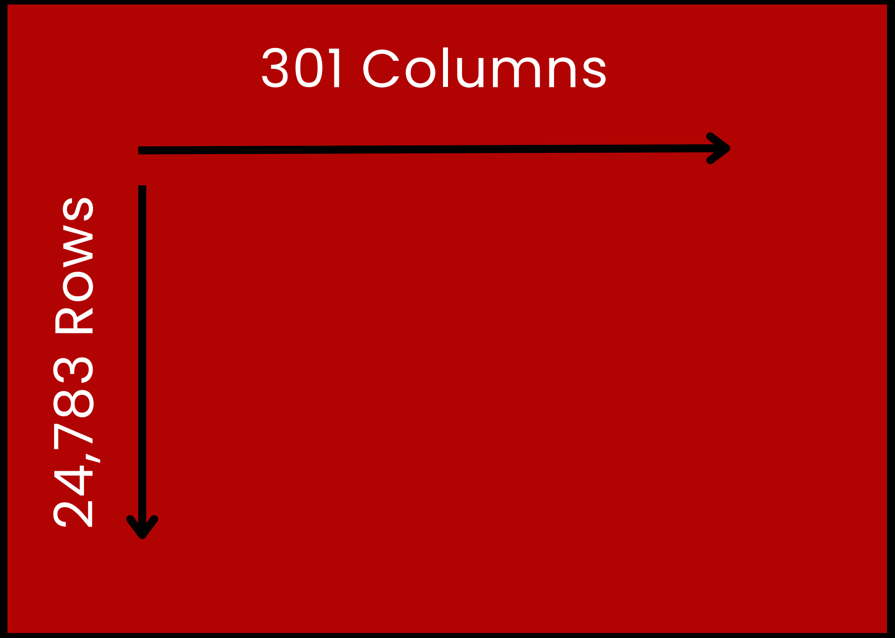
(4)

Created DF ready to feed to classification model

Similarity Check - 'Cheese'

```
mac 0.553615927696228
fries 0.5464779138565063
simpsons 0.5272508263587952
cookies 0.5202698707580566
fxx8217s 0.5170167088508606
httptcomy5ojyz8w9 0.51667541265
cornbread 0.5068199634552002
beans 0.5059381723403931
```

Features Matrix



DOC2VEC

{'vector_size': , 'window': , 'min_count': , 'epochs': , 'min_alpha': }

(1)
Initialized and trained various parameter combos

(2)
Compared the models using the wordsim535 file

MODEL	SPEARMAN	PEARSON
Default	0.0635	0.06988
Large Vector	0.07228	0.044242
Small Vector	0.04477	0.0053
Wide Window	0.0219	-0.0049

MODEL	SPEARMAN	PEARSON
High Min_Count	0.1	0.2218
Many Epochs	0.2419	0.2814
Skip Gram	0.031	-0.042
High LR	0.08383	-0.000608

DOC2VEC

(3)

Checked that model works

(4)

Created DF ready to feed to classification model

Vectorizer Check - 'did we win last night'

```
[-7.80629367e-02  2.78813485e-02  1.70882288e-02 -8.09044614e-02
  2.19768852e-01  6.27341568e-02 -1.16145434e-02 -1.89851701e-01
 -9.42868739e-02 -4.07065243e-01 -1.90426521e-02 -4.45404761e-02
 -3.32417578e-01 -3.75866354e-01 -7.64090195e-02 -7.22804293e-02
  1.05594993e-01  6.76372927e-03 -2.60740280e-01 -2.65800387e-01
 -6.38148412e-02 -1.96982965e-01  8.66274387e-02 -7.37726986e-02
  2.31051922e-01 -4.10025328e-01 -2.40647763e-01  1.23361856e-01
 -6.00629263e-02 -2.44405791e-01 -3.05665284e-01  3.37436825e-01
 -3.84516865e-01  9.32723209e-02 -5.28409779e-02  2.43669122e-01
  1.72409609e-01 -2.24393874e-01 -1.13676779e-01 -3.31023373e-02
  3.05172563e-01 -1.11420201e-02  7.45483711e-02 -1.41186282e-01
  3.17301273e-01  3.13091397e-01  1.47826657e-01 -4.48466912e-02
 -1.84792697e-01  2.62305468e-01 -2.56339103e-01 -7.83389583e-02
  9.64028761e-02  1.56431705e-01 -3.37400347e-01  2.52183855e-01
  2.54459113e-01  2.12153897e-01  5.83371446e-02 -1.68928489e-01
  1.44821033e-01 -1.63881510e-01  8.10753778e-02 -3.52491111e-01
  1.83684066e-01 -7.99813028e-03  2.42106840e-01  1.31862268e-01
 -8.58441144e-02 -3.29364866e-01 -1.92828953e-01  2.66785100e-02
  2.21485525e-01 -9.50041041e-02 -1.16204932e-01 -1.49085253e-01
 -2.31400505e-01  5.13406359e-02 -1.22458629e-01  7.20605627e-02
 -9.94458050e-02 -1.57721117e-01  3.36928070e-01  2.13073775e-01
  1.20948203e-01  1.37993527e-04  1.24631830e-01 -5.00963442e-03
 -1.90752879e-01 -4.85521704e-02  2.71933168e-01 -6.50784746e-02
 -3.56853694e-01  2.69032687e-01 -1.49701104e-01  1.14643984e-01
  7.78503437e-03  5.82882129e-02 -7.89572224e-02  4.14703488e-02
 -4.64011729e-02 -2.00024724e-01  1.14387579e-01 -1.38374045e-01
  8.05586129e-02 -1.01908863e-01 -9.09568518e-02 -9.98684466e-02
 -6.12260066e-02  9.75493789e-02 -1.25681624e-01 -3.40149850e-02
 -1.81139320e-01  3.46054852e-01 -2.35486016e-01 -1.74212798e-01
 -3.67778838e-01  2.98019852e-02  1.53608948e-01 -9.17350221e-03
 -2.75620759e-01 -4.98487763e-02  2.87676677e-02  2.84986407e-01
  2.12473899e-01 -9.16114673e-02  1.04152672e-02  3.10073197e-02
 -2.87891626e-02 -6.91682100e-02 -5.95165305e-02 -5.16365357e-02
  8.49698111e-02 -4.31333452e-01  8.29280019e-02  4.70479019e-02
  7.02981427e-02  1.12247318e-01  6.94945529e-02 -3.03062908e-02
 -1.73424054e-02  6.83447858e-03 -1.65004179e-01 -7.10849911e-02
  1.07664391e-01 -3.13256122e-02 -1.70064405e-01 -2.09511086e-01
 -1.67602807e-01 -9.39875245e-02  2.89091885e-01 -1.51906282e-01
 -4.02439028e-01 -1.18518323e-01 -8.25839192e-02  5.35181463e-02
 -1.28819495e-01 -3.16148341e-01  1.22427560e-01  3.43072087e-01
 -1.01012737e-01 -3.22604813e-02  1.26285344e-01 -1.17871590e-01
 -2.09820732e-01 -5.31798378e-02  7.99030438e-02 -2.07645550e-01
  1.62360236e-01  2.23988995e-01  1.99184462e-01  2.18941748e-01
 -2.24814057e-01 -3.42822187e-02 -7.06542060e-02 -1.97608232e-01
 -1.67684436e-01  4.10564654e-02  2.67814964e-01 -1.87278762e-01
 -9.85843837e-02  1.97201356e-01  1.35049447e-01 -2.95259863e-01
 -2.86880583e-01  2.80770883e-02  1.74776599e-01  2.75648445e-01
  5.10394454e-01  7.31143728e-02  3.05965990e-02 -1.42516881e-01
  4.45020152e-03  9.70849395e-02 -2.35259250e-01 -1.79055426e-02
 -8.02400336e-02  5.90872914e-02  2.62400299e-01  6.04637749e-02
 -1.40632451e-01 -6.15822636e-02  4.21631373e-02  8.35529789e-02
```

```
-3.67778838e-01  2.98019852e-02  1.53608948e-01 -9.17350221e-03
-2.75620759e-01 -4.98487763e-02  2.87676677e-02  2.84986407e-01
 2.12473899e-01 -9.16114673e-02  1.04152672e-02  3.10073197e-02
-2.87891626e-02 -6.91682100e-02 -5.95165305e-02 -5.16365357e-02
 8.49698111e-02 -4.31333452e-01  8.29280019e-02  4.70479019e-02
 7.02981427e-02  1.12247318e-01  6.94945529e-02 -3.03062908e-02
-1.73424054e-02  6.83447858e-03 -1.65004179e-01 -7.10849911e-02
 1.07664391e-01 -3.13256122e-02 -1.70064405e-01 -2.09511086e-01
-1.67602807e-01 -9.39875245e-02  2.89091885e-01 -1.51906282e-01
-4.02439028e-01 -1.18518323e-01 -8.25839192e-02  5.35181463e-02
-1.28819495e-01 -3.16148341e-01  1.22427560e-01  3.43072087e-01
-1.01012737e-01 -3.22604813e-02  1.26285344e-01 -1.17871590e-01
-2.09820732e-01 -5.31798378e-02  7.99030438e-02 -2.07645550e-01
 1.62360236e-01  2.23988995e-01  1.99184462e-01  2.18941748e-01
-2.24814057e-01 -3.42822187e-02 -7.06542060e-02 -1.97608232e-01
-1.67684436e-01  4.10564654e-02  2.67814964e-01 -1.87278762e-01
-9.85843837e-02  1.97201356e-01  1.35049447e-01 -2.95259863e-01
-2.86880583e-01  2.80770883e-02  1.74776599e-01  2.75648445e-01
 5.10394454e-01  7.31143728e-02  3.05965990e-02 -1.42516881e-01
 4.45020152e-03  9.70849395e-02 -2.35259250e-01 -1.79055426e-02
-8.02400336e-02  5.90872914e-02  2.62400299e-01  6.04637749e-02
-1.40632451e-01 -6.15822636e-02  4.21631373e-02  8.35529789e-02
-3.48628983e-02  2.24301323e-01 -1.86099745e-02  4.25264925e-01
-7.31043145e-03 -2.57901400e-01 -1.46450670e-02 -1.51160523e-01
 1.98623464e-01 -5.11469543e-02  1.11722164e-01 -1.29868105e-01
-1.12855650e-01 -3.82591963e-01 -2.85338461e-01 -7.76243433e-02
 1.86985165e-01 -1.57905132e-01 -1.35168955e-01  1.18401192e-01
-6.81886449e-02 -1.99350759e-01 -1.80066153e-01 -1.93537995e-01
-2.32020542e-01  1.19854808e-01  1.00712538e-01 -1.68097526e-01
 5.59359007e-02 -1.28843069e-01 -5.09458557e-02  1.28690209e-02
 7.50067970e-03 -5.52085005e-02 -3.46150994e-02  5.55276684e-03
 2.21074820e-01 -3.43885332e-01 -1.53688595e-01  3.45437914e-01
 1.20999329e-01 -1.96970135e-01  4.96768951e-02  1.03273503e-01
-3.21168862e-02  1.86989844e-01 -1.76572934e-01  1.42369837e-01
-1.10005900e-01 -4.41701636e-02 -1.05714193e-02 -7.75635540e-02
 2.09789276e-01  1.81059644e-01 -2.67466277e-01 -7.78876478e-03
 2.35936403e-01  2.26122037e-01 -5.85799031e-02 -1.85731500e-01
-1.18735559e-01 -3.08864206e-01  6.32860735e-02  1.17192619e-01
-1.97841465e-01 -5.88222733e-03 -6.76722080e-02 -5.16662933e-02
-7.04727471e-02 -9.24598351e-02  4.19788122e-01  2.83528715e-01
-5.57829738e-02 -1.27148479e-02  9.80159640e-02  5.59232896e-03
 2.29454115e-01  8.32784250e-02  4.13225144e-02  1.48005232e-01
-4.74235602e-02 -2.26602200e-02 -1.94713309e-01  1.27988890e-01
 1.44695625e-01 -1.14919443e-03 -2.11396161e-02  2.01256126e-01
 8.79413169e-03  4.35675606e-02  8.87849629e-02 -2.11588308e-01
-3.25385071e-02  3.47046852e-02 -3.27882590e-03  4.64742146e-02]
```

DOC2VEC

(3)

Checked that model works

(4)

Created DF ready to feed to classification model

Similar Docs - 'yo yo yo'

```
Similar Documents (Doc2Vec):
Document ID: 23201, Similarity: 0.8120245933532715
Document ID: 6386, Similarity: 0.792421102523803 7
Document ID: 90, Similarity: 0.789602935314178 5
Document ID: 605, Similarity: 0.775378525257110 6
Document ID: 16750, Similarity: 0.7490344047546387
Document ID: 21253, Similarity: 0.7422741651535034
Document ID: 17678, Similarity: 0.7387194037437439
Document ID: 8783, Similarity: 0.7385451793670654
Document ID: 9410, Similarity: 0.7347460389137268
Document ID: 6183, Similarity: 0.7254233956336975
```

DATA FRAME SUMMARY

TF-IDF
36,301 columns
24,783 rows

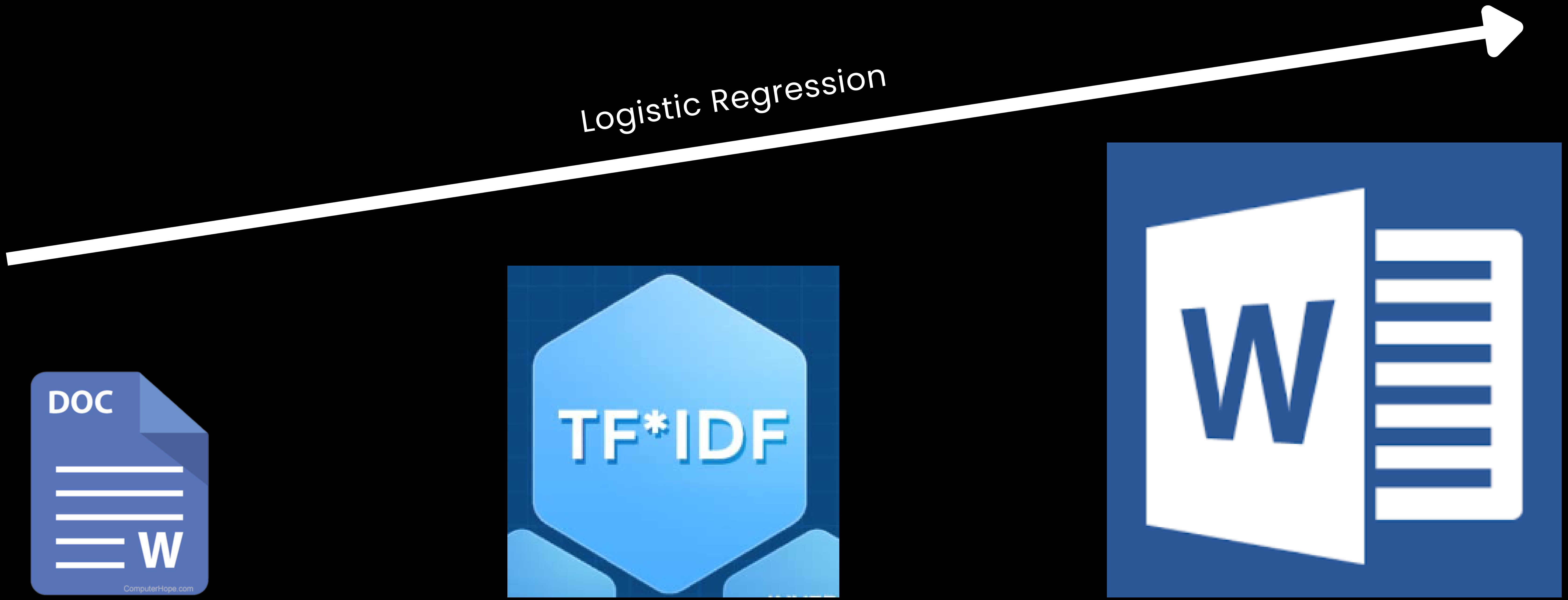
Word2Vec
301 columns
24,783 rows

Doc2Vec
301 columns
24,783 rows

CLASSIFICATION MODELS

Logistic Regression	86%
Random Forrest	86%
XG_Boost	84%
LDA	74%

CLASSIFICATION RESULTS



PROJECT SUMMARY

THE COMBINATION OF LOGISTIC REGRESSION WITH WORD2VEC VECTORIZATION
88% ACCURACY


HATE SPEECH