MINOR PROJECT

Profiling Hate Speech Spreaders on Twitter

- Under supervision of Dr. Jyoti Prakash

PRESENTED BY:

- Rakshita Jain(1806136)
- Devanshi Goel(1806180)
- Prashant Sahu(1806171)

6th Semester, Cse, NIT Patna

Problem

- The importance of hate speech detection research cannot be overemphasised.
- Now, more than ever, with the current inflammatory political climate and discourse all around the world and minorities in various locations demanding for equality and equity, we cannot allow additional bias to be introduced into their lives through artificial intelligence.
- The problem of hate speech detection is one yet to be solved even to an acceptable level.
- It would be counter-productive if all the research efforts are not focused and channeled towards a better tomorrow by building on top one another.
- So we were motivated to go back to a root of the problem: the data

Project Motive

We aim at identifying possible hate speech spreaders on Twitter as a first step towards preventing hate speech from being propagated among online users.

After having several aspects of author profiling in social media from 2013 to 2020 (fake news spreaders, bot detection, age and gender, also together with personality, gender and language variety, and gender from a multimodality perspective) addressed, we aim at investigating if it is possbile to discriminate authors that have shared some hate speech in the past from those that, to the best of our knowledge, have never done it.

This is a PAN @ CLEF 2021 event task

DATASET

DATASET NAME

PAN21-Profiling-Hate-Speech-Spreaders-in-Twitter

SOURCE

pan.webis.de

DATA

- train data (200 authors)
  - Split ratio - train:0.67, test: 0.33

FORMAT OF GIVEN DATASET

- XMl files for various tweet ids'
- Each XML file contains 100 tweets of a particular author
- Two languages - English & Spanish
- Label file in text format containing the labels for training dataset

MODIFICATION DONE TO DATASET

- Converted XML and text files to csv
- Merged all the csv files of different tweet ids' into one file (combined.csv)
- Merged all the tweets of one tweet id into a single entry
- Added label as a column in the training dataset.

Preprocessing

1. Removed duplicated tweets
2. Removed hash symbol and hardcode mentions to username
3. Removed stop words and punctuations using nltk
4. Convert Emojis to words
5. Stemming using Porter Stemmer
6. Tokenization using tweet tokenizer

DATA AFTER PREPROCESSING

	id	tweet	label
0	06ct0t68y1acizh9eow3g5rhancrppr8	courteney cox recreates classic friends scene ...	1
1	071nxc49ihpd0jlfmvn2lghtayy3b5n9	amber smith kandy halloween return haunted man...	0
2	09py5qescynpnnckmzueqzr2y49moh1o	rachel bilson asked point blank shes dating ni...	0
3	0dwovd7nj6yg9m795ng2c629me0ccmrh	public relations officer zone police command g...	0
4	0ibi364m7i7l01xi4xqafyathrmrrnll	know started know began fighting intensifies g...	1
...	...	...	...
295	zuelpgcp4186rxrhifbslyrdfhhaxxt8	court overturns rescue ship captain conviction...	0
296	zurv8xodgwcle1guhjai6n1i4cw4lc8r	antonio burgos indignant twitter message bimba...	1
297	zw2pjht6tf3ymkfbfbm83zcjxfuumzal	home studio essentials beginners starttechnolo...	1
298	zwfesexkazacsz78p8g1h6ockrcvoypf	quiz well know old panic disco let us know sco...	0
299	zzsafm5u4tzk6k0ba500tlggn7iw8v03	bridget marquardt seen hugh hefner years bridg...	1

300 rows × 3 columns

FEATURES

	00	000	01	04	10	100	10poundsme	10user	11	12	1200	13	137k	14	1400	140k	14user	15	150one	16	17	176k	1861	1887confusion	1892confusion	19	191
0	0	0	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	
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	
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	
4	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	2	0	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
129	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
130	1	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
131	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
132	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
133	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Count vectorizer features

FEATURE EXTRACTION

1

TD-IDF(n-grams)

2

Count vectorizor

3

Word embedding

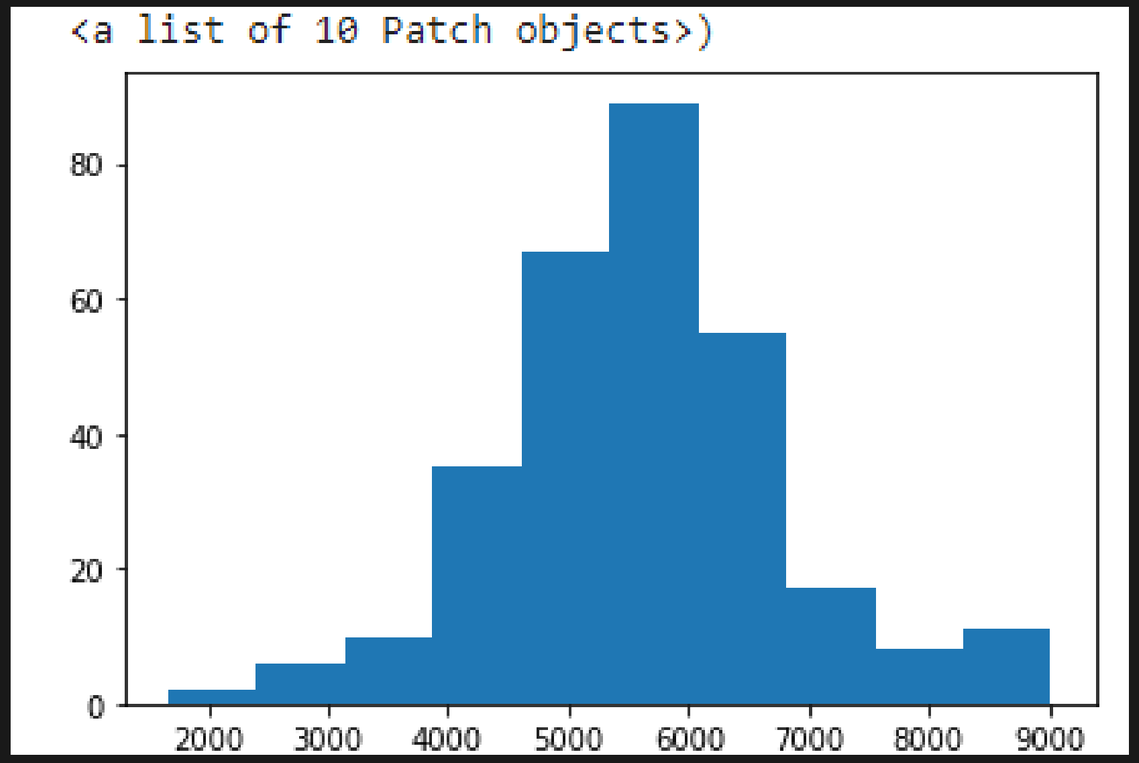
4

One Hot encoding

DATA VISUALISATION

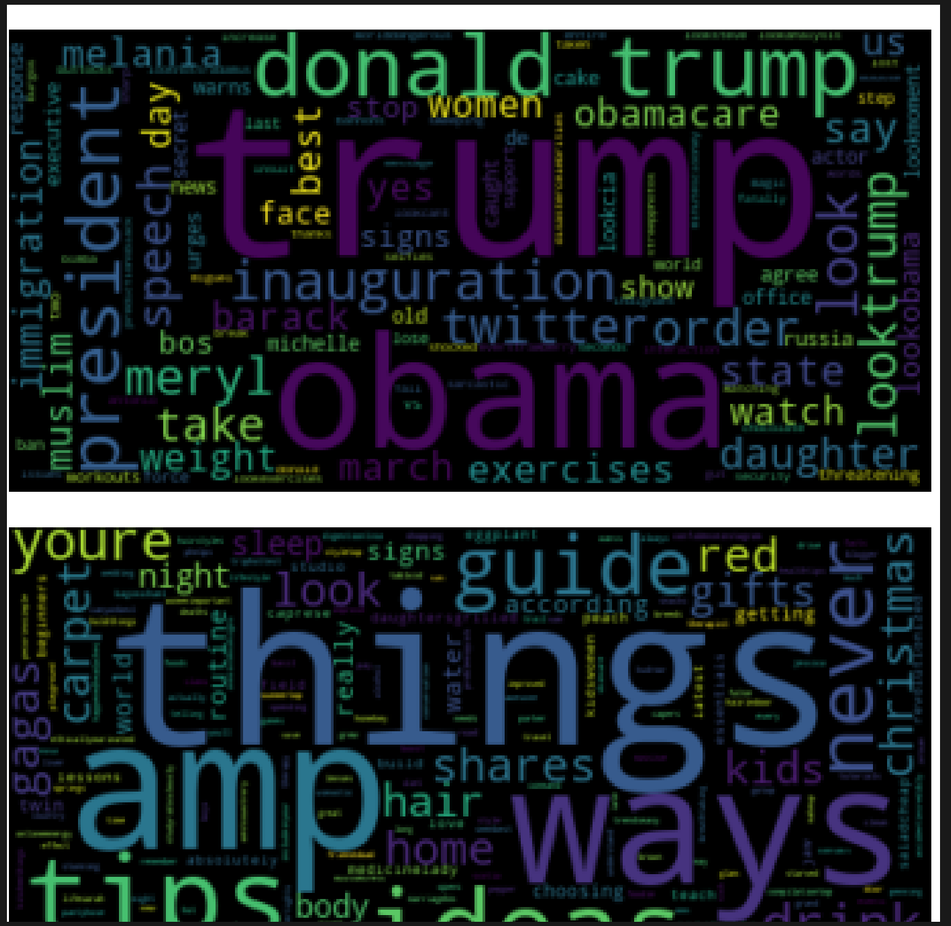
1

Visualising combined tweet length using matplotlib



2

Word cloud



MODELS IMPLEMENTED

ML models

1

Naive Bayes (using tfidf & count vectorizer)

2

K-Neighbours (using tfidf & count vectorizer)

3

Logistic Regression

3

Linear SVC SVM

Deep Learning Models

4

LSTM (using word embedding)

6

Bi-LSTM

5

BERT

English

N-grams range - (1, 1)

Naive Bayes - Count vectorizer

	Precision	Recall	F1 - Score	Support
0	0.79	0.72	0.75	36
1	0.7	0.77	0.73	30
weighted avg	0.75	0.74	0.74	66
Accuracy	0.742			
Alpha	0.2			

Naive Bayes - TF-IDF

	Precision	Recall	F1 - Score	Support
0	0.72	0.72	0.72	36
1	0.67	0.67	0.67	30
weighted avg	0.7	0.7	0.7	66
Accuracy	0.696			
Alpha	0.1			

K Neighbours - Count Vectorizer

	Precision	Recall	F1 - Score	Support
0	0.55	0.72	0.63	36
1	0.47	0.3	0.37	30
weighted avg	0.52	0.53	0.51	66
Accuracy	0.53			
N	2			

English

N-grams range - (1, 3)

Naive Bayes - Count vectorizer

	Precision	Recall	F1 - Score	Support
0	0.76	0.53	0.62	36
1	0.59	0.8	0.68	30
weighted avg	0.68	0.65	0.65	66
Accuracy	0.65			
Alpha	0			

Naive Bayes - TF-IDF

	Precision	Recall	F1 - Score	Support
0	0.72	0.58	0.65	36
1	0.59	0.73	0.66	30
weighted avg	0.67	0.65	0.65	66
Accuracy	0.65			
Alpha	0			

K Neighbours - Count Vectorizer

	Precision	Recall	F1 - Score	Support
0	0.62	0.58	0.6	36
1	0.53	0.57	0.55	30
weighted avg	0.58	0.58	0.58	66
Accuracy	0.575			
N	9			

OBTAINED ACCURACY

OBTAINED ACCURACY

English				
N-grams range - (1, 1)				
K Neighbours - TF-IDF				
	Precision	Recall	F1 - Score	Support
0	0.73	0.61	0.67	36
1	0.61	0.73	0.67	30
weighted avg	0.68	0.67	0.67	66
Accuracy	0.66			
N	6			
Logistic Regression - Count vect				
	Precision	Recall	F1 - Score	Support
0	0.67	0.61	0.64	36
1	0.58	0.63	0.6	30
weighted avg	0.63	0.62	0.62	66
Accuracy	0.6212			
Logistic Regression - TF-IDF				
	Precision	Recall	F1 - Score	Support
0	0.75	0.5	0.6	36
1	0.57	0.8	0.67	30
weighted avg	0.67	0.64	0.63	66
Accuracy	0.636			

English				
N-grams range - (1, 3)				
K Neighbours - TF-IDF				
	Precision	Recall	F1 - Score	Support
0	0.67	0.67	0.67	36
1	0.6	0.6	0.6	30
weighted avg	0.64	0.64	0.64	66
Accuracy	0.636			
N	6			
Logistic Regression - Count vect				
	Precision	Recall	F1 - Score	Support
0	0.76	0.61	0.68	36
1	0.62	0.77	0.69	30
weighted avg	0.7	0.68	0.68	66
Accuracy	0.6818			
Logistic Regression - TF-IDF				
	Precision	Recall	F1 - Score	Support
0	0.88	0.39	0.54	36
1	0.56	0.93	0.7	30
weighted avg	0.73	0.64	0.61	66
Accuracy	0.636			

	English			
	N-grams range - (1, 1)			
	SVC SVM - TF-IDF			
	Precision	Recall	F1 - Score	Support
0	0.75	0.5	0.6	36
1	0.57	0.8	0.67	30
weighted avg	0.67	0.64	0.63	66
Accuracy	0.636			

	English			
	N-grams range - (1, 3)			
	Linear SVC SVM - TF IDF			
	Precision	Recall	F1 - Score	Support
0	0.88	0.39	0.54	36
1	0.56	0.93	0.7	30
weighted avg	0.73	0.64	0.61	66
Accuracy	0.636			

Bi LSTM	
Accuracy	54 %
Epochs	15
LSTM	
Accuracy	44%
Epochs	15
BERT	
Accuracy	53%
Epochs	15

OBTAINED ACCURACY

Highest - 74 %

Spanish				
N-grams range - (1, 1)				
Naive Bayes - Count Vectoriser				
	Precision	Recall	F1 - Score	Support
0	0.78	0.78	0.78	32
1	0.79	0.79	0.79	34
weighted avg	0.79	0.79	0.79	66
Accuracy	0.787			
Alpha	0.2			
Naive Bayes - TF-IDF				
	Precision	Recall	F1 - Score	Support
0	0.79	0.81	0.8	32
1	0.82	0.79	0.81	34
weighted avg	0.8	0.8	0.8	66
Accuracy	0.8			
Alpha	0.1			
K Neighbours - Count Vectorizer				
	Precision	Recall	F1 - Score	Support
0	0.74	0.81	0.78	32
1	0.81	0.74	0.77	34
weighted avg	0.78	0.77	0.77	66
Accuracy	0.77			
N	6			

Spanish				
N-grams range - (1, 3)				
Naive Bayes - Count Vectoriser				
	Precision	Recall	F1 - Score	Support
0	0.86	0.75	0.8	32
1	0.79	0.88	0.83	34
weighted avg	0.82	0.82	0.82	66
Accuracy	0.81			
Alpha	0.1			
Naive Bayes - TF-IDF				
	Precision	Recall	F1 - Score	Support
0	0.86	0.75	0.8	32
1	0.79	0.88	0.83	34
weighted avg	0.82	0.82	0.82	66
Accuracy	0.82			
Alpha	0			
K Neighbours - Count Vectorizer				
	Precision	Recall	F1 - Score	Support
0	0.71	0.78	0.75	32
1	0.77	0.71	0.74	34
weighted avg	0.75	0.74	0.74	66
Accuracy	0.74			
N	7			

OBTAINED ACCURACY

Spanish

N-grams range - (1, 1)

K Neighbours - TF-IDF

	Precision	Recall	F1 - Score	Support
0	0.75	0.56	0.64	32
1	0.67	0.82	0.74	34
weighted avg	0.71	0.7	0.69	66
Accuracy	0.69			
N	2			

Logistic Regression - Count vect

	Precision	Recall	F1 - Score	Support
0	0.77	0.84	0.81	32
1	0.84	0.76	0.8	34
weighted avg	0.81	0.8	0.8	66
Accuracy	0.8			

Logistic Regression - TF-IDF

	Precision	Recall	F1 - Score	Support
0	0.79	0.72	0.75	32
1	0.76	0.82	0.79	34
weighted avg	0.77	0.77	0.77	66
Accuracy	0.77			

Spanish

N-grams range - (1, 3)

K Neighbours - TF-IDF

	Precision	Recall	F1 - Score	Support
0	0.8	0.25	0.38	32
1	0.57	0.94	0.71	34
weighted avg	0.68	0.61	0.55	66
Accuracy	0.6			
N	6			

Logistic Regression - Count vect

	Precision	Recall	F1 - Score	Support
0	0.76	0.81	0.79	32
1	0.81	0.76	0.79	34
weighted avg	0.79	0.79	0.79	66
Accuracy	0.79			

Logistic Regression - TF-IDF

	Precision	Recall	F1 - Score	Support
0	0.84	0.66	0.74	32
1	0.73	0.88	0.8	34
weighted avg	0.78	0.77	0.77	66
Accuracy	0.77			

OBTAINED ACCURACY

	Spanish			
	N-grams range - (1, 1)			
Linear SVC SVM - Count vectoriser				
	Precision	Recall	F1 - Score	Support
0	1	0.06	0.12	32
1	0.53	1	0.69	34
weighted avg	0.76	0.55	0.41	66
Accuracy	0.54			
Linear SVC SVM - TF-IDF				
	Precision	Recall	F1 - Score	Support
0	0.79	0.72	0.75	32
1	0.76	0.82	0.79	34
weighted avg	0.77	0.77	0.77	66
Accuracy	0.77			

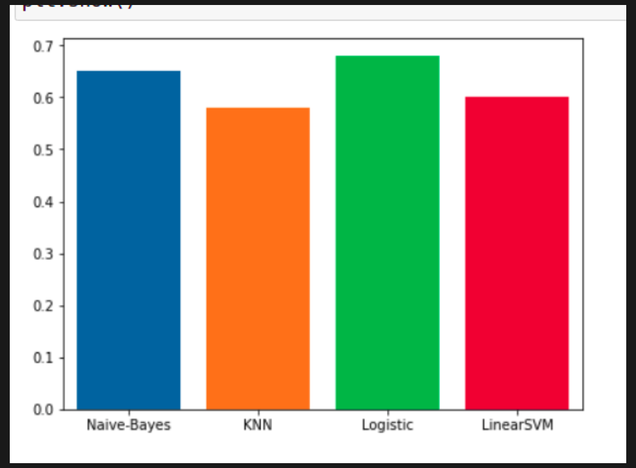
	Spanish			
	N-grams range - (1, 3)			
Linear SVC SVM - Count vectoriser				
	Precision	Recall	F1 - Score	Support
0	1	0.06	0.12	32
1	0.53	1	0.69	34
weighted avg	0.76	0.55	0.41	66
Accuracy	0.54			
Linear SVC SVM - TF-IDF				
	Precision	Recall	F1 - Score	Support
0	0.84	0.66	0.74	32
1	0.73	0.88	0.8	34
weighted avg	0.78	0.77	0.77	66
Accuracy	0.77			

OBTAINED ACCURACY

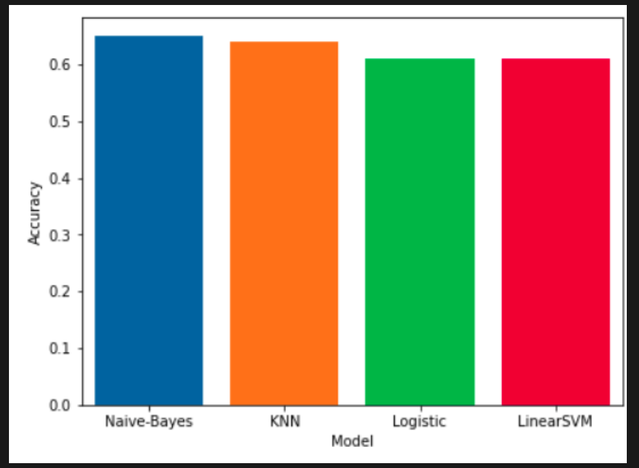
Bi LSTM		LSTM		BERT	
Accuracy	48.5 %	Accuracy	46.9%	Accuracy	56.06%
Epochs	15	Epochs	15	Epochs	15

Highest - 82 %

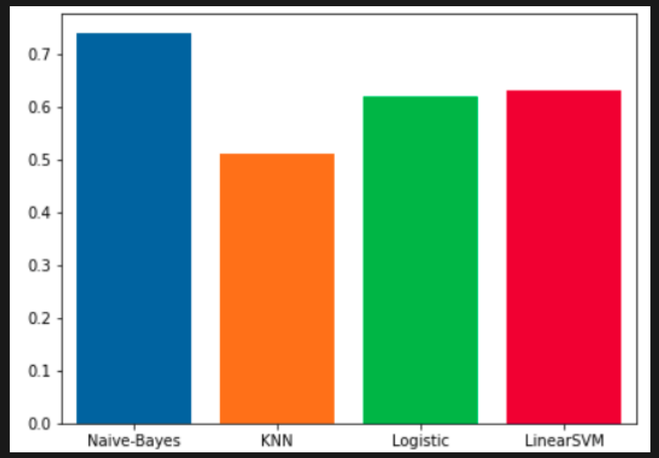
English



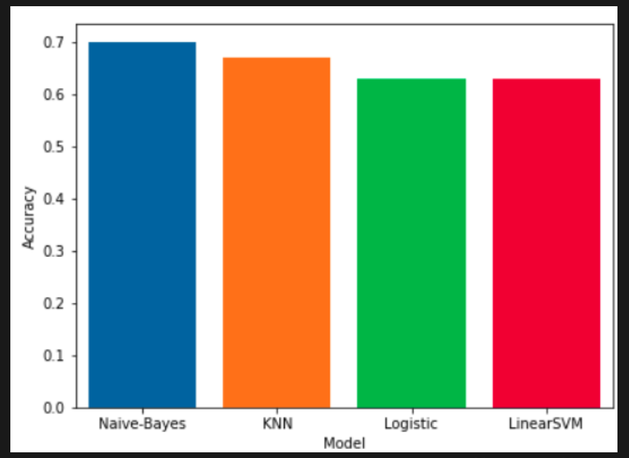
n-grams(1,3), count vectoriser



n-grams(1,3), tf-idf

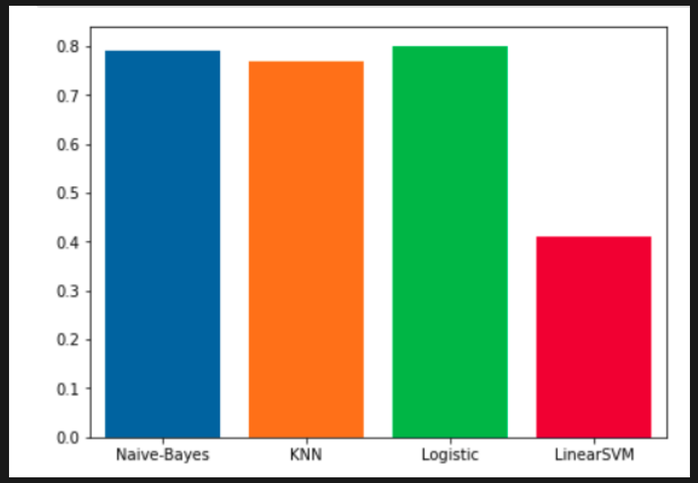


n-grams(1,1), count vectoriser

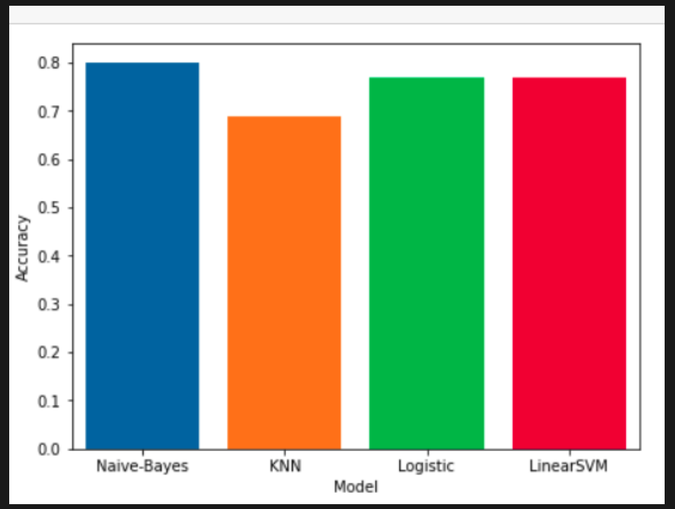


n-grams(1,1), tf-idf

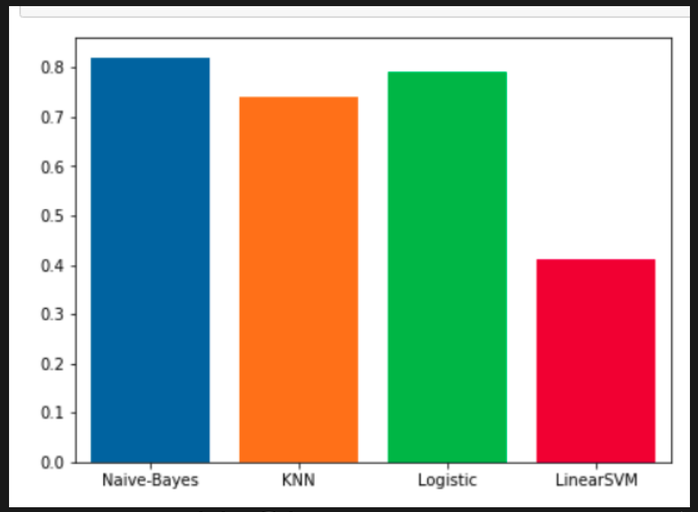
Spanish



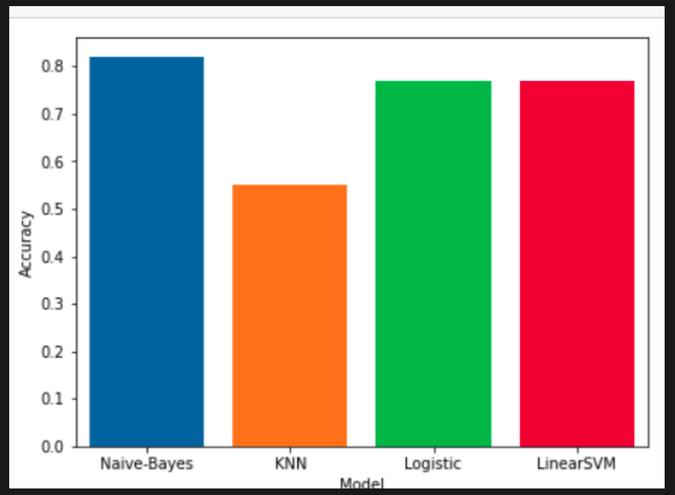
n-grams(1,1), count vectoriser



n-grams(1,1), tf-idf



n-grams(1,3), count vectoriser



n-grams(1,3), tf-idf

CONCLUSION

For English Dataset	For Spanish Dataset
Model - Naive Bayes	Model - Naive Bayes
F1-Score - 0.74	**F1-Score - 0.82**

- We observed that Deep Learning Models did not perform well for the given problem.

For English Data, we found that Naive Bayes Classifier performed best

- for the feature Count vectoriser
- with n-gram range (1,1)

For Spanish Data, we found that Naive Bayes Classifier performed best:

- for both the features Count vectoriser and TF-IDF
- with n-gram range (1,3)

References

- [1] Samuel Caetano da Silva, Thiago Castro Ferreira, Ricelli Moreira Silva Ramos, Ivandre Paraboni (2020). Data-driven and psycholinguistics motivated approaches to hate speech detection. Computación y Sistemas, 24(3): 1179–1188
- [2] Stiven Zimmerman, Udo Kruschwitz, Cris Fox (2018). Improving hate speech detection with deep learning ensembles. In Proc. of the Eleventh Int. Conf. on Language Resources and Evaluation (LREC 2018)
- [3] Simona Frenda, Bilal Ghanem, Manuel Montes-y Gomez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. Journal of Intelligent & Fuzzy Systems, 36(5):4743–4752.
- [4] Francisco Rangel, Anastasia Giachanou, Bilal Ghanem, Paolo Rosso. Overview of the 8th Author Profiling Task at PAN 2020: Profiling Fake News Spreaders on Twitter. In: L. Cappellato, C. Eickhoff, N. Ferro, and A. Névéol (eds.) CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org, vol. 2696
- [5] Francisco Rangel and Paolo Rosso. Overview of the 7th Author Profiling Task at PAN 2019: Bots and Gender Profiling in Twitter. In: L. Cappellato, N. Ferro, D. E. Losada and H. Müller (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings.CEUR-WS.org, vol. 2380
- https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3478