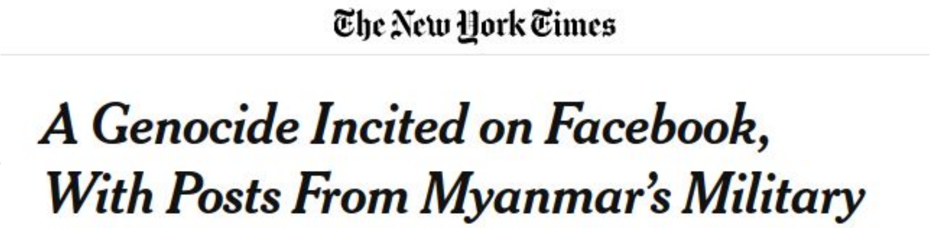
Detección de Hate Speech en Twitter

MOTIVACIÓN

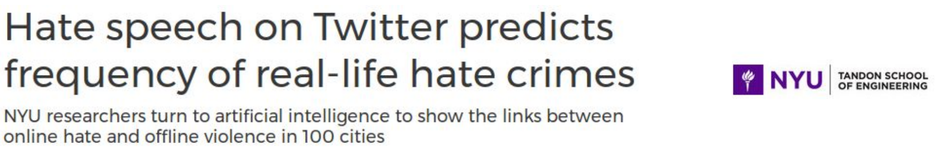
El discurso del odio en redes sociales conduce a la normalización de la discriminación, la intolerancia y a actitudes y comportamientos hostiles.

MOTIVACIÓN



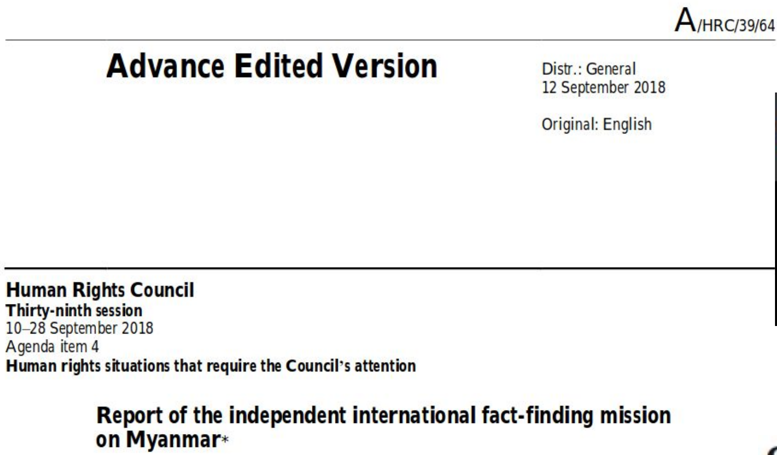
The New York Times

A Genocide Incited on Facebook, With Posts From Myanmar's Military



Hate speech on Twitter predicts frequency of real-life hate crimes

NYU researchers turn to artificial intelligence to show the links between online hate and offline violence in 100 cities



A/HRC/39/64

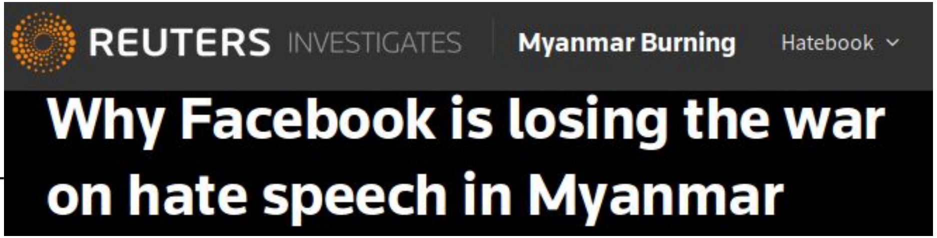
Advance Edited Version

Distr.: General
12 September 2018

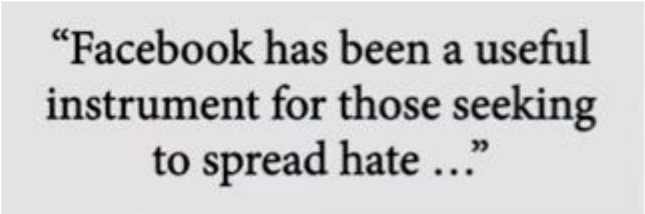
Original: English

Human Rights Council
Thirty-ninth session
10–28 September 2018
Agenda item 4
Human rights situations that require the Council's attention

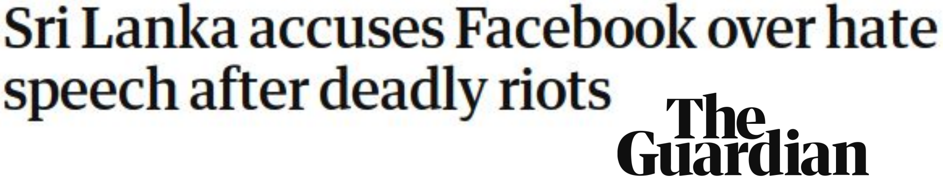
Report of the independent international fact-finding mission on Myanmar*



Why Facebook is losing the war on hate speech in Myanmar



"Facebook has been a useful instrument for those seeking to spread hate ..."

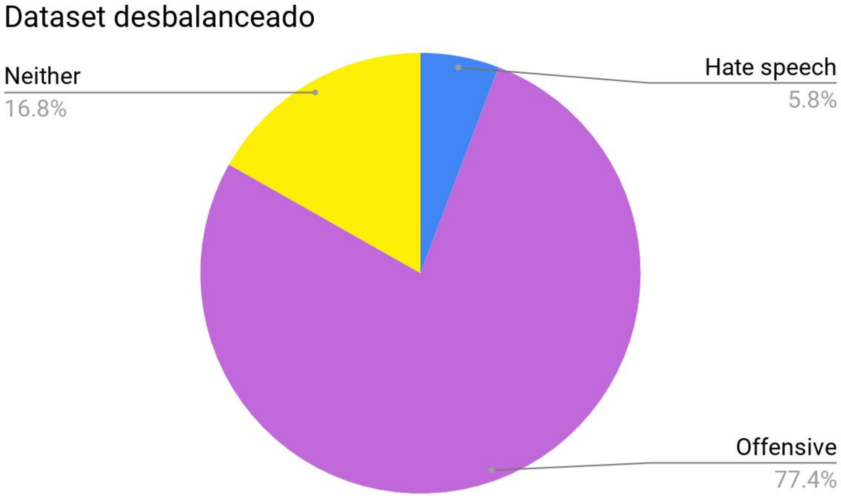


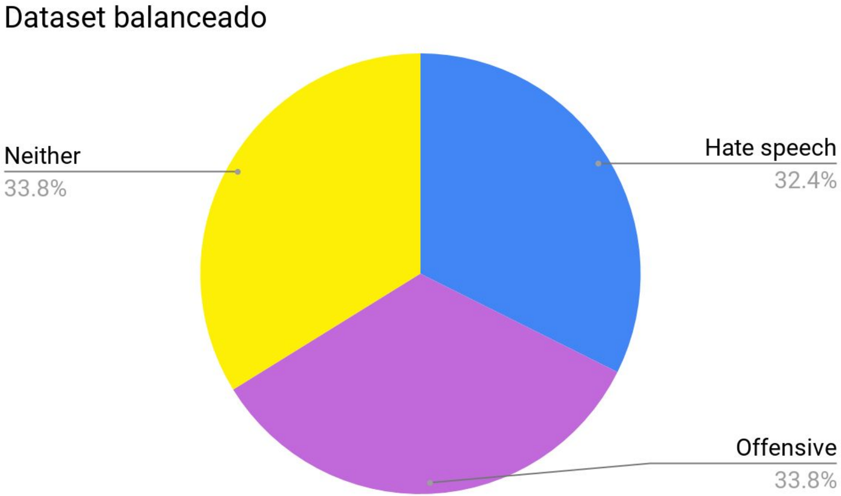
Sri Lanka accuses Facebook over hate speech after deadly riots

The Guardian

Dataset

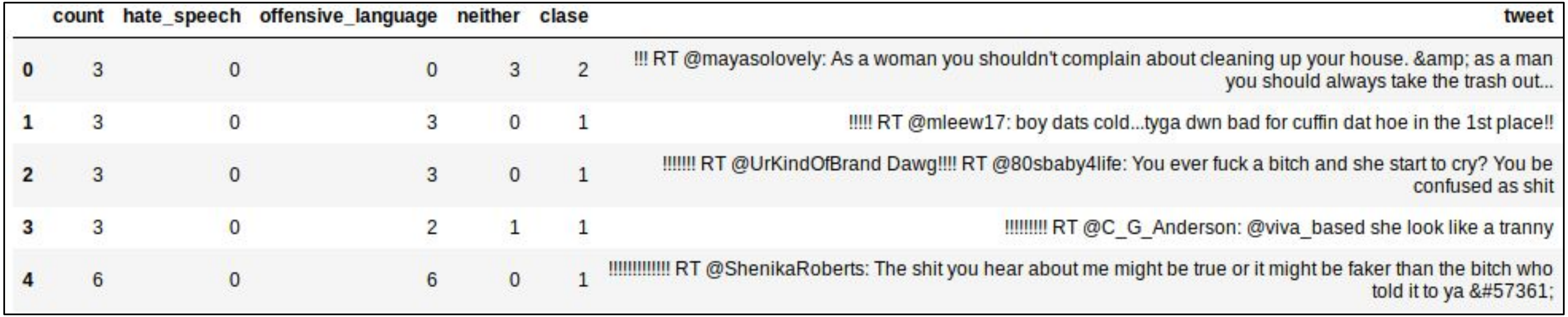
HATEBASE


Dataset desbalanceado

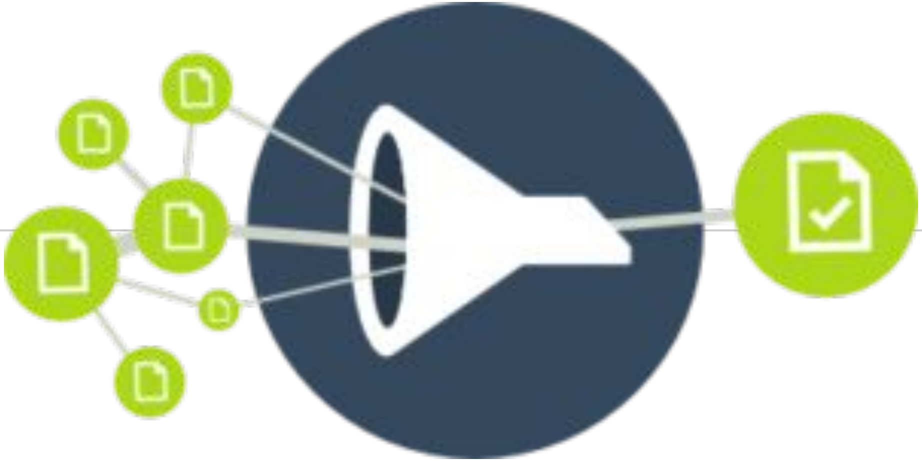

Dataset balanceado

Hate speech: *Cualquier comunicación que desacredite a una persona o un grupo en función de alguna característica, como raza, color, origen étnico, género, orientación sexual, nacionalidad, religión u otra característica.*

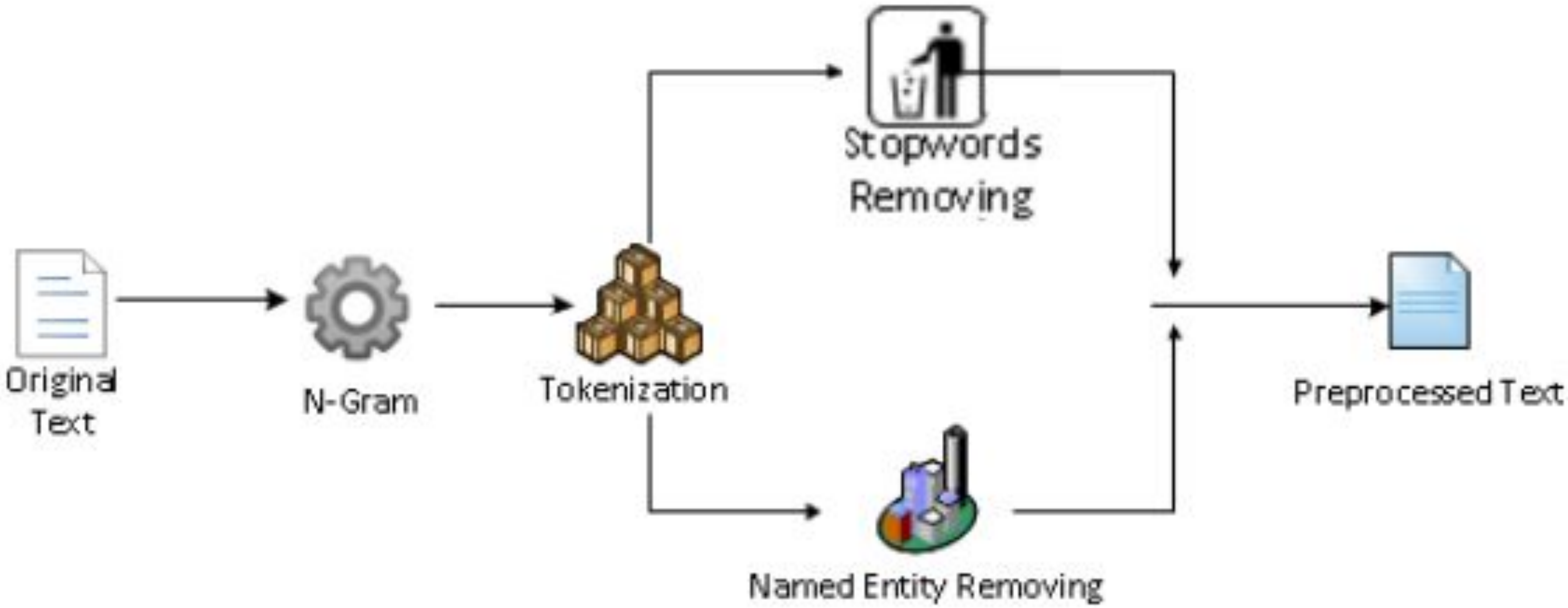
Nockleby (2002)



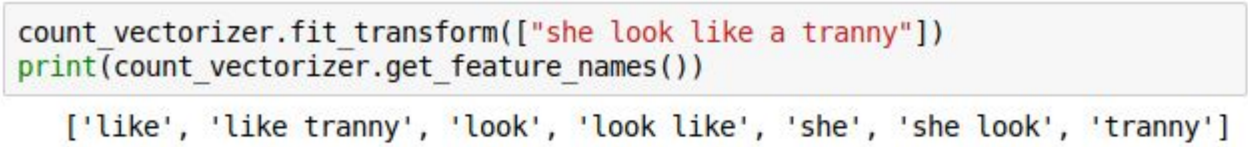
	count	hate_speech	offensive_language	neither	clase	tweet
0	3	0	0	3	2	!!! RT @mayasolovely: As a woman you shouldn't complain about cleaning up your house. &amp; as a man you should always take the trash out...
1	3	0	3	0	1	!!!!! RT @mleew17: boy dats cold...tyga dwn bad for cuffin dat hoe in the 1st place!!
2	3	0	3	0	1	!!!!!! RT @UrKindOfBrand Dawg!!!! RT @80sbaby4life: You ever fuck a bitch and she start to cry? You be confused as shit
3	3	0	2	1	1	!!!!!!!!! RT @C_G_Anderson: @viva_based she look like a tranny
4	6	0	6	0	1	!!!!!!!!!!!!! RT @ShenikaRoberts: The shit you hear about me might be true or it might be faker than the bitch who told it to ya &#57361;



Preprocesamiento


Original Text
N-Gram
Tokenization
Stopwords Removing
Named Entity Removing
Preprocessed Text

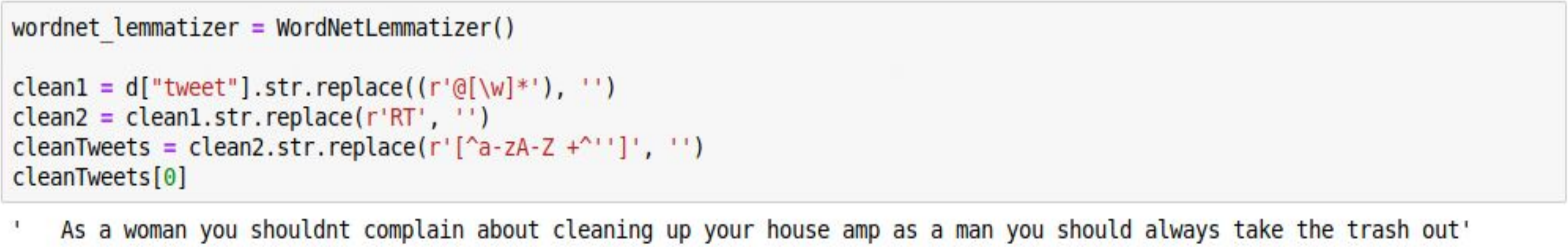
N-gram and tokenization



```
count_vectorizer.fit_transform(["she look like a tranny"])
print(count_vectorizer.get_feature_names())
```

```
['like', 'like tranny', 'look', 'look like', 'she', 'she look', 'tranny']
```

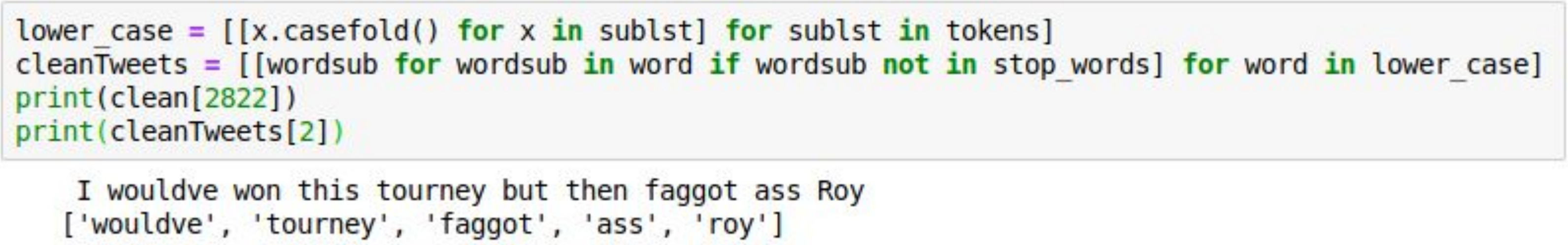
Named entity removing



```
wordnet_lemmatizer = WordNetLemmatizer()

clean1 = d["tweet"].str.replace((r'@[\w]*'), '')
clean2 = clean1.str.replace(r'RT', '')
cleanTweets = clean2.str.replace(r'[^a-zA-Z +^'']', '')
cleanTweets[0]
```

```
'   As a woman you shouldnt complain about cleaning up your house amp as a man you should always take the trash out'
```

Stopwords removing

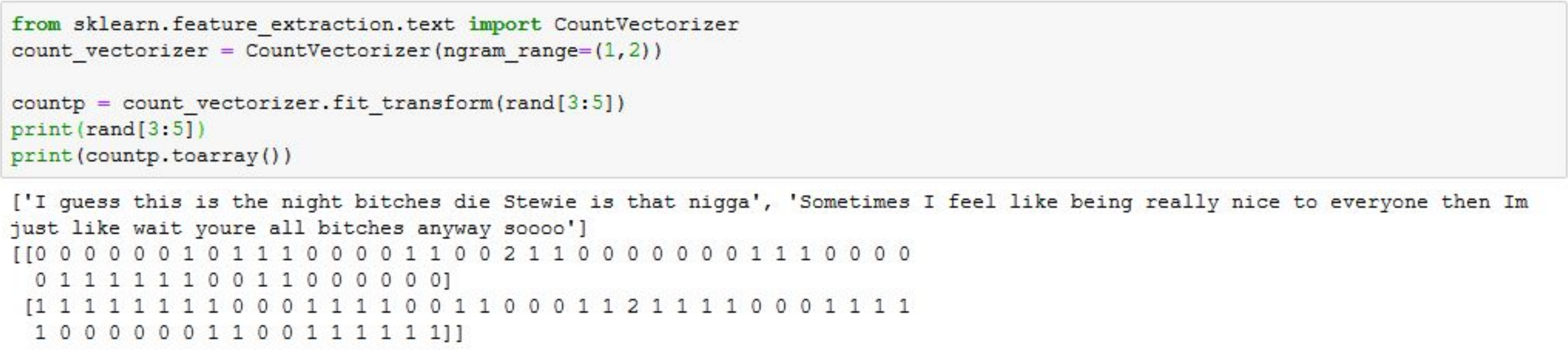


```
lower_case = [[x.casefold() for x in sublst] for sublst in tokens]
cleanTweets = [[wordsub for wordsub in word if wordsub not in stop_words] for word in lower_case]
print(clean[2822])
print(cleanTweets[2])
```

```
 I wouldve won this tourney but then faggot ass Roy
['wouldve', 'tourney', 'faggot', 'ass', 'roy']
```

Transformación de los datos

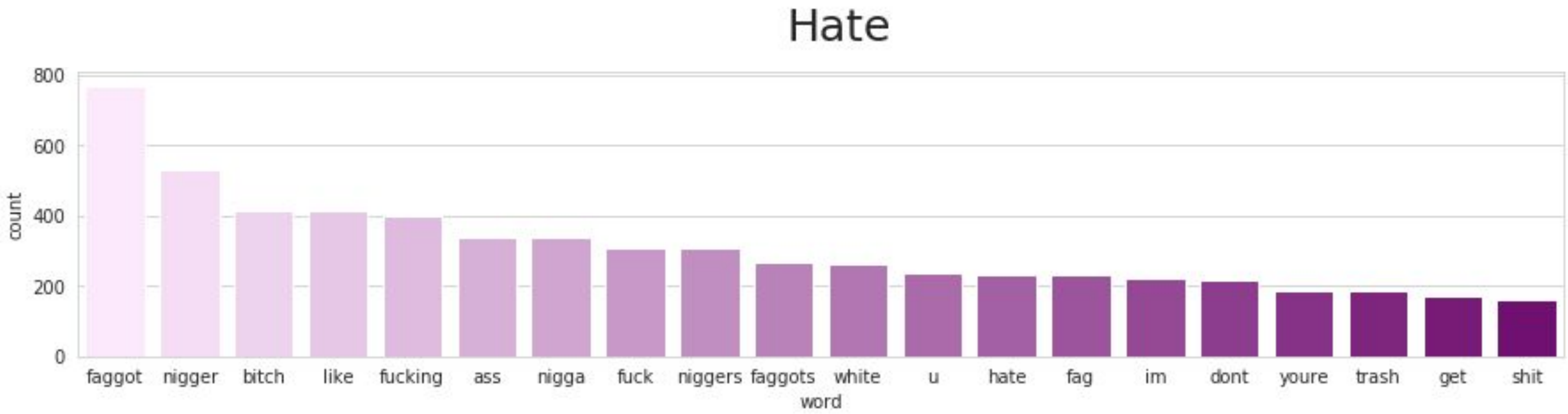
CountVectorizer



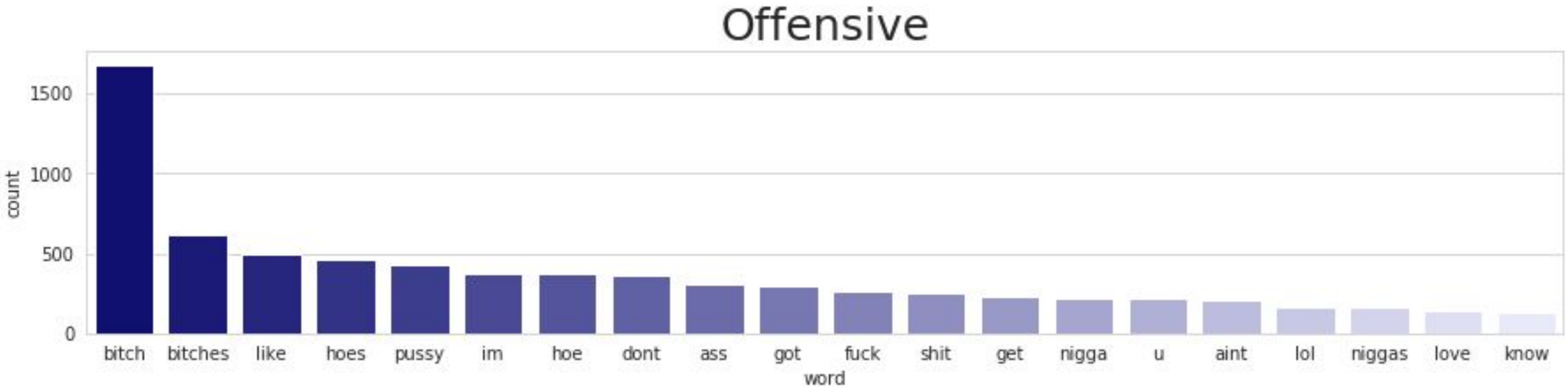
```
from sklearn.feature_extraction.text import CountVectorizer
count_vectorizer = CountVectorizer(ngram_range=(1,2))

countp = count_vectorizer.fit_transform(rand[3:5])
print(rand[3:5])
print(countp.toarray())
```

```
['I guess this is the night bitches die Stewie is that nigga', 'Sometimes I feel like being really nice to everyone then Im
just like wait youre all bitches anyway soooo']
[[0 0 0 0 0 0 1 0 1 1 1 0 0 0 0 1 1 0 0 2 1 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0
  0 1 1 1 1 1 1 0 0 1 1 0 0 0 0 0 0]
 [1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 0 0 1 1 0 0 0 1 1 2 1 1 1 1 0 0 0 1 1 1 1
  1 0 0 0 0 0 0 1 1 0 0 1 1 1 1 1 1]]
```

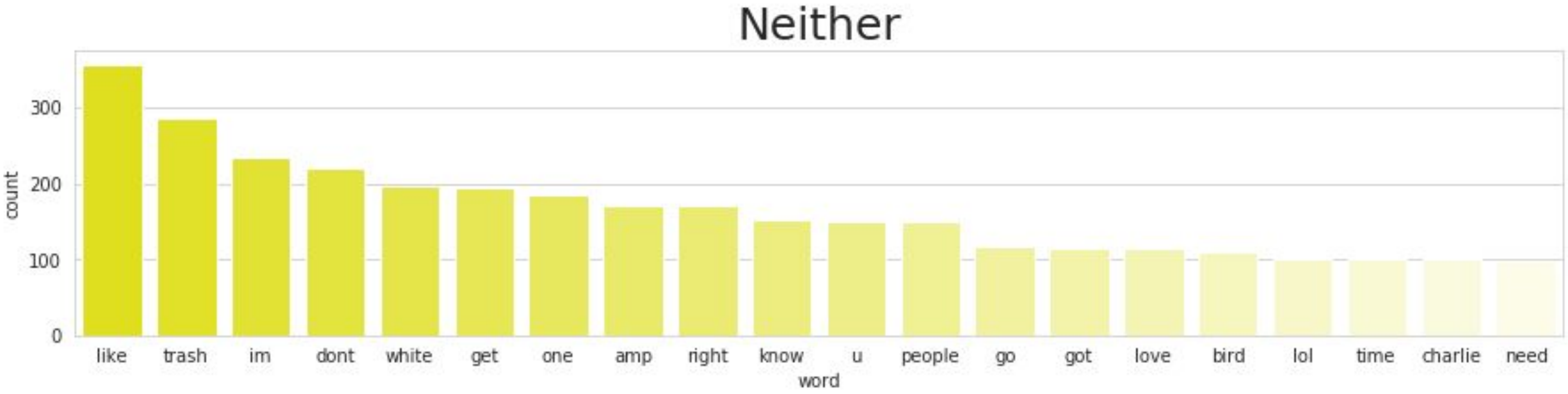
Exploración del dataset

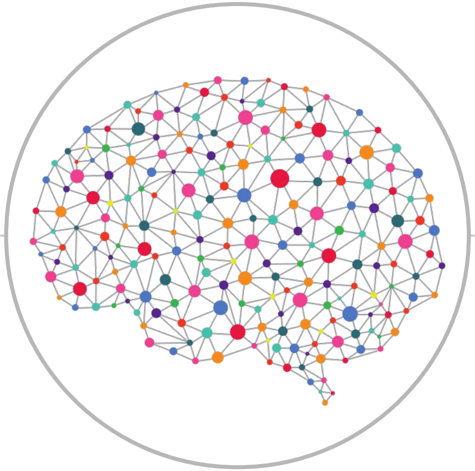


Exploración del dataset



Exploración del dataset





Clasificación

1 Cross validation

MultinomialNB

alpha= 0.5
cv= KFold(10, shuffle=True)

85.83% accuracy (balanceado)
74.25% accuracy (desbalanceado)

RandomForestClassifier

cv= StratifiedKFold(3, shuffle=True)

92.21 % accuracy (balanceado)
83.95% accuracy (desbalanceado)

2 One-Vs-the-Rest

SVC

probability=True
kernel= ‘linear’

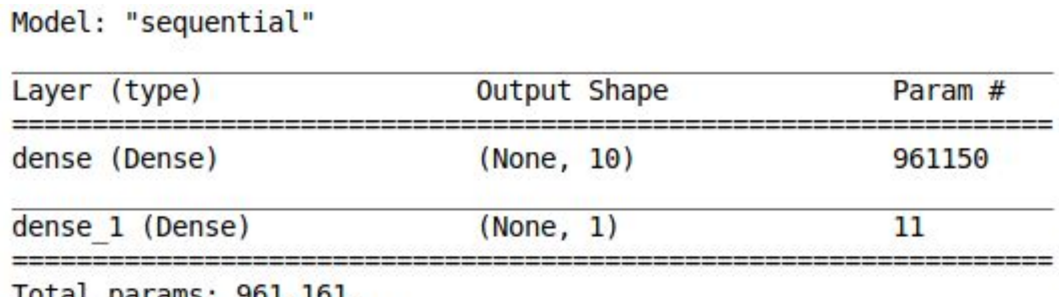
86.28% accuracy (balanceado)
88.84% accuracy (desbalanceado)

MultinomialNB

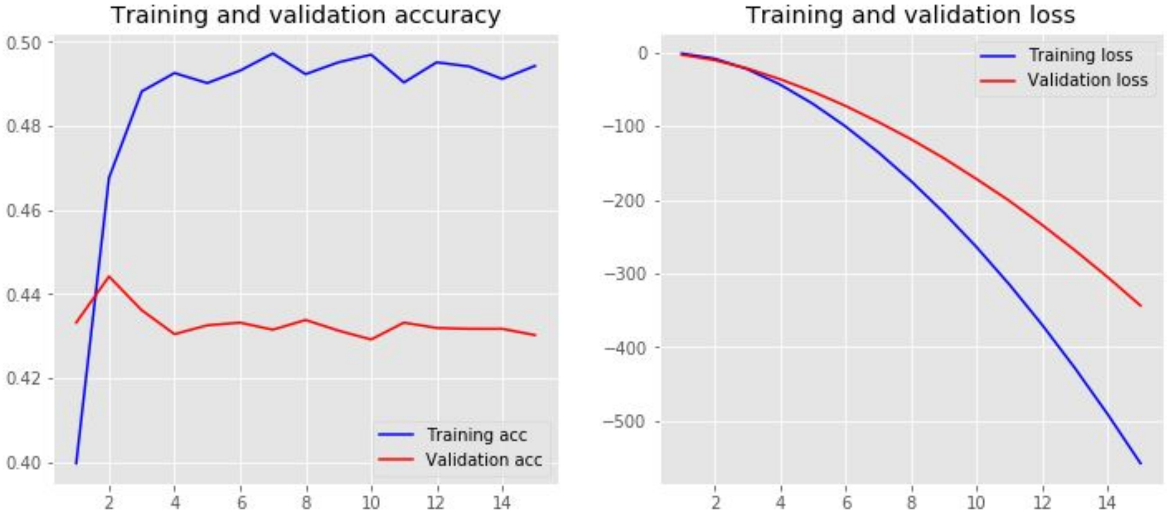
alpha= 0.5

83.11 % accuracy (balanceado)
79.01% accuracy (desbalanceado)

3 Red neuronal



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 10)                961150
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 11
=================================================================
Total params: 961,161
Trainable params: 961,161
Non-trainable params: 0
```



Librerias

matplotlib

scikit learn

NLTK

TensorFlow + Keras

Gracias!

¿Preguntas?

Repositorio Github

- https://github.com/vgoyenechec/Hatespeech-en-twitter