upgrade_first_project

March 23, 2023

1 More Attempts on my First NLP Model

In this notebook I want to provide some new attempts made in my first NLP project.

I performed more EDA and feature engineering by trying different encoding methods, so the results may be different from my first attempt, where, for example, some duplicated values were kept.

It will need further improvement for sure.

I am doing it on the lessons provided by the following link: https://www.youtube.com/@codebasics

Original notebook: https://github.com/flaviobrienza/Fake_News_Classifier/blob/main/first_nlp_project.ipynb

2 Importing Libraries

[1]:
```
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

sns.set_style("darkgrid")
matplotlib.rcParams['font.size'] = 14
matplotlib.rcParams['figure.figsize'] = (15, 5)
matplotlib.rcParams['figure.facecolor'] = '#00000000'

import warnings
warnings.simplefilter(action='ignore')

import spacy
```

3 Importing Data

[34]:
```
df =  pd.read_csv('./fake-or-real-news/fake_or_real_news.csv').
 ↪drop(columns='Unnamed: 0')

df
```

[34]:
```
                                                  title  \
0                          You Can Smell Hillary's Fear
1      Watch The Exact Moment Paul Ryan Committed Pol…
2            Kerry to go to Paris in gesture of sympathy
3      Bernie supporters on Twitter erupt in anger ag…
4       The Battle of New York: Why This Primary Matters
…                                                    …
6330   State Department says it can't find emails fro…
6331   The 'P' in PBS Should Stand for 'Plutocratic' …
6332   Anti-Trump Protesters Are Tools of the Oligarc…
6333   In Ethiopia, Obama seeks progress on peace, se…
6334   Jeb Bush Is Suddenly Attacking Trump. Here's W…

                                                   text label
0      Daniel Greenfield, a Shillman Journalism Fello…  FAKE
1      Google Pinterest Digg Linkedin Reddit Stumbleu…  FAKE
2      U.S. Secretary of State John F. Kerry said Mon…  REAL
3      - Kaydee King (@KaydeeKing) November 9, 2016 T…  FAKE
4      It's primary day in New York and front-runners…  REAL
…                                                    …     …
6330   The State Department told the Republican Natio…  REAL
6331   The 'P' in PBS Should Stand for 'Plutocratic' …  FAKE
6332    Anti-Trump Protesters Are Tools of the Oligar…  FAKE
6333   ADDIS ABABA, Ethiopia -President Obama convene…  REAL
6334   Jeb Bush Is Suddenly Attacking Trump. Here's W…  REAL

[6335 rows x 3 columns]
```

4 EDA and Feature Engineering

[35]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6335 entries, 0 to 6334
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   title   6335 non-null   object
 1   text    6335 non-null   object
 2   label   6335 non-null   object
dtypes: object(3)
memory usage: 148.6+ KB
```

There are no missing values.

Dropping duplicated values.

[36]:
```
df[df.text.duplicated(keep=False)]
```

[36]:
```
                                                  title  \
12    Strong Solar Storm, Tech Risks Today | S0 News…
14                      Trump takes on Cruz, but lightly
25    Anti-Trump forces seek last-ditch delegate revolt
30                        GOP insiders: Carly crushed it
35    Mike Pence Drapes Shawl Over Immodest Lady Jus…
…                                                   …
6227  ISIS uses an industrial dough kneader to kill …
6233  North Korea Threatens ‘Sacred’ Nuclear War Aga…
6250           Activists bristle at Clinton fundraising
6270      Inside Bernie Sanders’ unorthodox debate prep
6328  Radio Derb Is On The Air-Leonardo And Brazil’s…

                                                   text label
12    Click Here To Learn More About Alexandra's Per…  FAKE
14    Killing Obama administration rules, dismantlin…  REAL
25    Washington (CNN) The faction of the GOP that i…  REAL
30    On this day in 1973, J. Fred Buzhardt, a lawye…  REAL
35    Trump Raises Concern Over Members Of Urban Com…  FAKE
…                                                   …     …
6227  Email \nISIS barbarians used an industrial dou…  FAKE
6233  Email \nNorth Korea’s Foreign Ministry slammed…  FAKE
6250  A verdict in 2017 could have sweeping conseque…  REAL
6270  Killing Obama administration rules, dismantlin…  REAL
6328                                                    FAKE

[344 rows x 3 columns]
```

[37]:
```
df = df.drop_duplicates('text')
df.shape
```

[37]:
```
(6060, 3)
```

[38]:
```
df[df.duplicated('title', keep=False)]
```

[38]:
```
                                                  title  \
174   Syrian War Report - November 2, 2016: ISIS and…
271   The Fix Is In: NBC Affiliate Accidentally Post…
423   US abstains from UN vote calling for end to Cu…
451   Get Ready For Civil Unrest: Survey Finds That …
529   Meteor, space junk, rocket? Mysterious flash h…
…                                                   …
6214  Tony Blair suggests a second referendum to rev…
6231  Schools All Over America Are Closing On Electi…
6303  The Deceptive Nature of Hillary Clinton is Rig…
6307  US abstains from UN vote calling for end to Cu…
6329  Assange claims ‘crazed’ Clinton campaign tried…
```

```
                                                    text label
174   Trump Whistles His Dogs < > South Front Analys…  FAKE
271   NBC affiliate WRCB TV in Chattanooga, Tennesse…  FAKE
423   US abstains from UN vote calling for end to Cu…  FAKE
451   in: Protestors & Activists , Special Interests…  FAKE
529   Meteor, space junk, rocket? Mysterious flash h…  FAKE
…                                                   …    …
6214  Tony Blair suggests a second referendum to rev…  FAKE
6231  in: Politics , Sleuth Journal , Special Intere…  FAKE
6303  Posted by David Risselada \nMuch to the surpri…  FAKE
6307  US abstains from UN vote calling for end to Cu…  FAKE
6329  Julian Assange has claimed the Hillary Clinton…  FAKE

[95 rows x 3 columns]
```

[39]:
```
df = df.drop_duplicates('title')
df.shape
```

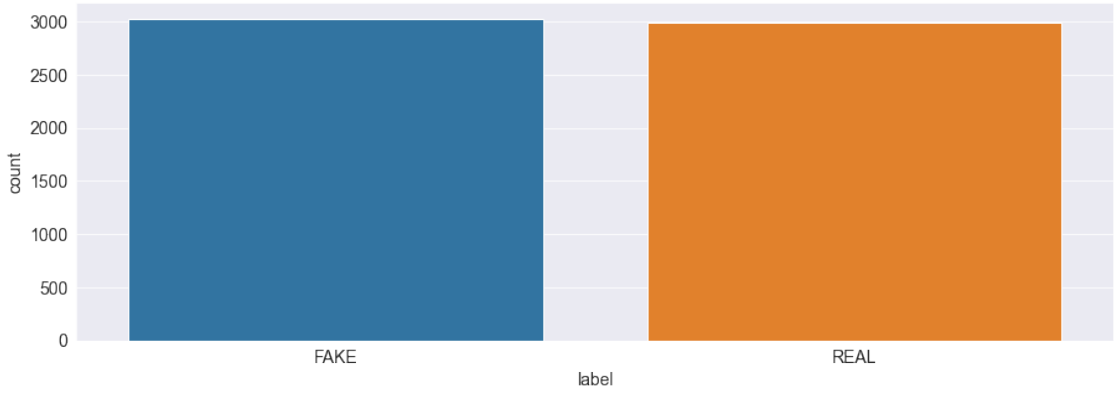
[39]:
```
(6011, 3)
```

Checking the class balance.

[43]:
```
df.label.value_counts()
```

[43]:
```
FAKE    3026
REAL    2985
Name: label, dtype: int64
```

[44]:
```
sns.countplot(data=df, x='label');
```



The dataset is almost perfectly balanced.

Checking the average length of the text for fake and real news.

[45]:
```
df['text_length'] = df.text.apply(lambda x: len(x))
df
```

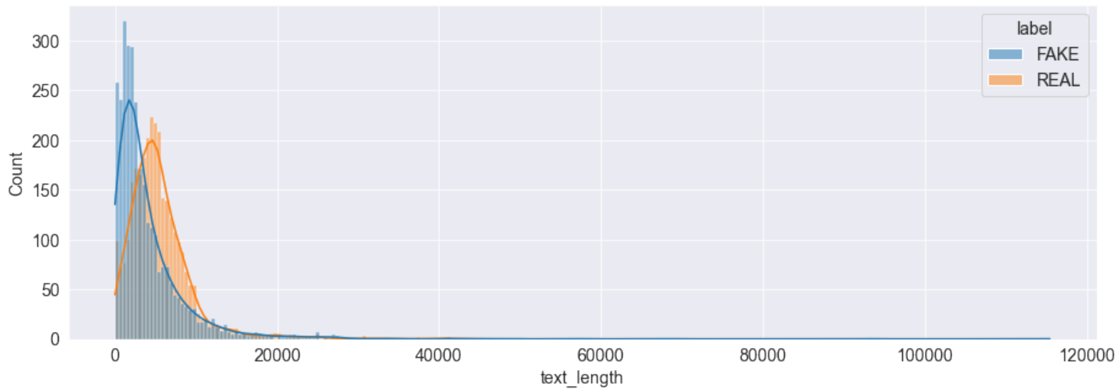
[45]:
```
                                                  title  \
0                           You Can Smell Hillary's Fear
1       Watch The Exact Moment Paul Ryan Committed Pol…
2             Kerry to go to Paris in gesture of sympathy
3       Bernie supporters on Twitter erupt in anger ag…
4        The Battle of New York: Why This Primary Matters
…                                                     …
6330    State Department says it can't find emails fro…
6331    The 'P' in PBS Should Stand for 'Plutocratic' …
6332    Anti-Trump Protesters Are Tools of the Oligarc…
6333    In Ethiopia, Obama seeks progress on peace, se…
6334    Jeb Bush Is Suddenly Attacking Trump. Here's W…

                                                   text label  text_length
0       Daniel Greenfield, a Shillman Journalism Fello…  FAKE         7518
1       Google Pinterest Digg Linkedin Reddit Stumbleu…  FAKE         2646
2       U.S. Secretary of State John F. Kerry said Mon…  REAL         2543
3       - Kaydee King (@KaydeeKing) November 9, 2016 T…  FAKE         2660
4       It's primary day in New York and front-runners…  REAL         1840
…                                                     …     …           …
6330    The State Department told the Republican Natio…  REAL         4076
6331    The 'P' in PBS Should Stand for 'Plutocratic' …  FAKE        14323
6332     Anti-Trump Protesters Are Tools of the Oligar…  FAKE        11974
6333    ADDIS ABABA, Ethiopia -President Obama convene…  REAL         6991
6334    Jeb Bush Is Suddenly Attacking Trump. Here's W…  REAL         4818

[6011 rows x 4 columns]
```

[46]:
```
df[df.label=='FAKE'].text_length.describe(), df[df.label=='REAL'].text_length.
↪describe()
```

[46]:
```
(count      3026.000000
 mean       4179.931593
 std        5742.954547
 min           1.000000
 25%        1312.500000
 50%        2597.000000
 75%        5061.000000
 max      115372.000000
 Name: text_length, dtype: float64,
 count     2985.000000
 mean      5582.642211
 std       4293.736327
 min         43.000000
```

```
25%        3063.000000
50%        4841.000000
75%        6971.000000
max       44039.000000
Name: text_length, dtype: float64)
```

[47]:
```
sns.histplot(data=df, x='text_length', hue='label', kde=True);
```



Real news seem to be longer than fake ones.
Is it the same for the title?

[48]:
```
df['title_length'] = df.title.apply(lambda x: len(x))
df
```

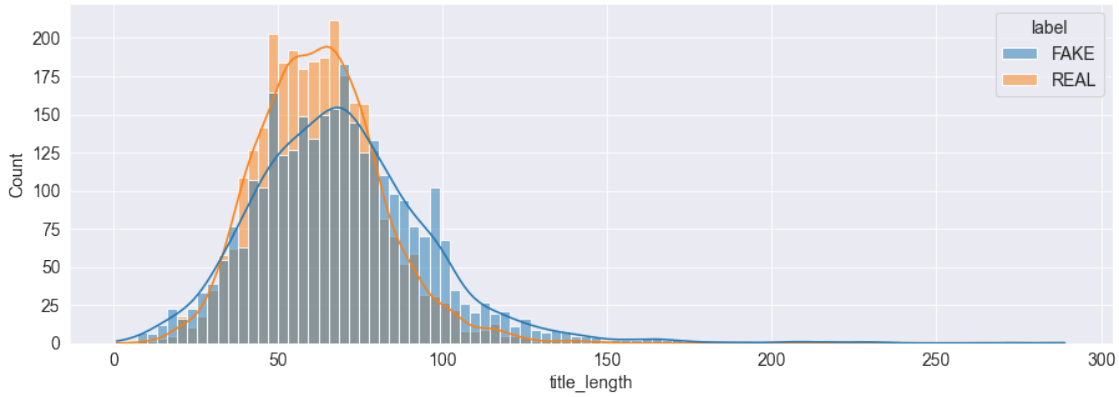
[48]:
```
                                                     title  \
0                             You Can Smell Hillary’s Fear
1       Watch The Exact Moment Paul Ryan Committed Pol…
2             Kerry to go to Paris in gesture of sympathy
3       Bernie supporters on Twitter erupt in anger ag…
4        The Battle of New York: Why This Primary Matters
…                                                      …
6330    State Department says it can't find emails fro…
6331    The ‘P’ in PBS Should Stand for ‘Plutocratic’ …
6332    Anti-Trump Protesters Are Tools of the Oligarc…
6333    In Ethiopia, Obama seeks progress on peace, se…
6334    Jeb Bush Is Suddenly Attacking Trump. Here's W…

                                                     text label  text_length  \
0       Daniel Greenfield, a Shillman Journalism Fello…  FAKE         7518
1       Google Pinterest Digg Linkedin Reddit Stumbleu…  FAKE         2646
2       U.S. Secretary of State John F. Kerry said Mon…  REAL         2543
3       — Kaydee King (@KaydeeKing) November 9, 2016 T…  FAKE         2660
4       It's primary day in New York and front-runners…  REAL         1840
```

```
…                                                     …   …           …
6330  The State Department told the Republican Natio…  REAL         4076
6331  The ‘P’ in PBS Should Stand for ‘Plutocratic’ …  FAKE        14323
6332   Anti-Trump Protesters Are Tools of the Oligar…  FAKE        11974
6333  ADDIS ABABA, Ethiopia -President Obama convene…  REAL         6991
6334  Jeb Bush Is Suddenly Attacking Trump. Here's W…  REAL         4818

      title_length
0               28
1               85
2               43
3               84
4               48
…              …
6330            69
6331            59
6332            66
6333            67
6334            61

[6011 rows x 5 columns]
```

[49]:
```
df[df.label=='FAKE'].title_length.describe(), df[df.label=='REAL'].title_length.
↪describe()
```

[49]:
```
(count    3026.000000
 mean       69.240251
 std        27.163359
 min         1.000000
 25%        51.000000
 50%        68.000000
 75%        84.000000
 max       289.000000
 Name: title_length, dtype: float64,
 count    2985.000000
 mean       62.627806
 std        18.661022
 min        10.000000
 25%        50.000000
 50%        62.000000
 75%        74.000000
 max       148.000000
 Name: title_length, dtype: float64)
```

[50]:
```
sns.histplot(data=df, x='title_length', hue='label', kde=True);
```



There is not much difference between the titles' lenght.

Changing the "label" feature into 0 and 1, where:

- **REAL** –> 0
- **FAKE** –> 1

[51]:
```
df['label'] = df.label.map({'REAL':0, 'FAKE':1})
df.head()
```

[51]:
```
                                               title  \
0                       You Can Smell Hillary's Fear
1  Watch The Exact Moment Paul Ryan Committed Pol…
2        Kerry to go to Paris in gesture of sympathy
3  Bernie supporters on Twitter erupt in anger ag…
4   The Battle of New York: Why This Primary Matters

                                                text  label  text_length  \
0  Daniel Greenfield, a Shillman Journalism Fello…      1         7518
1  Google Pinterest Digg Linkedin Reddit Stumbleu…      1         2646
2  U.S. Secretary of State John F. Kerry said Mon…      0         2543
3  - Kaydee King (@KaydeeKing) November 9, 2016 T…      1         2660
4  It's primary day in New York and front-runners…      0         1840

   title_length
0            28
1            85
2            43
3            84
4            48
```

Importing spacy model.

[52]:
```
nlp = spacy.load('en_core_web_sm')
```

A first attempt to improve the result will be to remove punctuation, stop words and applying lemmatization to the sentences.

[54]:
```
def lemmit(row):
    doc = nlp(row['text'])
    lemm_list = []
    for word in doc:
        if not word.is_punct and not word.is_stop:
            lemm_list.append(word.lemma_)
    return ' '.join(lemm_list)
```

[55]:
```
df['lemmit_text'] = df.apply(lemmit, axis=1)
```

Saving the preprocessed df.

[56]:
```
df.to_csv('preprocessed_fake_news.csv', index=None)
```

Showing the results.

[57]:
```
df
```

[57]:

```
                                                  title  \
0                          You Can Smell Hillary's Fear
1      Watch The Exact Moment Paul Ryan Committed Pol…
2            Kerry to go to Paris in gesture of sympathy
3      Bernie supporters on Twitter erupt in anger ag…
4       The Battle of New York: Why This Primary Matters
…                                                    …
6330   State Department says it can't find emails fro…
6331   The 'P' in PBS Should Stand for 'Plutocratic' …
6332   Anti-Trump Protesters Are Tools of the Oligarc…
6333   In Ethiopia, Obama seeks progress on peace, se…
6334   Jeb Bush Is Suddenly Attacking Trump. Here's W…

                                                   text  label  text_length  \
0      Daniel Greenfield, a Shillman Journalism Fello…      1         7518
1      Google Pinterest Digg Linkedin Reddit Stumbleu…      1         2646
2      U.S. Secretary of State John F. Kerry said Mon…      0         2543
3      - Kaydee King (@KaydeeKing) November 9, 2016 T…      1         2660
4      It's primary day in New York and front-runners…      0         1840
…                                                    …    …           …
6330   The State Department told the Republican Natio…      0         4076
6331   The 'P' in PBS Should Stand for 'Plutocratic' …      1        14323
6332    Anti-Trump Protesters Are Tools of the Oligar…      1        11974
6333   ADDIS ABABA, Ethiopia -President Obama convene…      0         6991
6334   Jeb Bush Is Suddenly Attacking Trump. Here's W…      0         4818

      title_length                                        lemmit_text
0               28  Daniel Greenfield Shillman Journalism Fellow F…
```

```
1               85  Google Pinterest Digg Linkedin Reddit Stumbleu…
2               43  U.S. Secretary State John F. Kerry say Monday …
3               84  Kaydee King @KaydeeKing November 9 2016 lesson…
4               48  primary day New York runner Hillary Clinton Do…
…              …                                                 …
6330            69  State Department tell Republican National Comm…
6331            59  p PBS stand plutocratic Pentagon post Oct 27 2…
6332            66    Anti Trump Protesters Tools Oligarchy  refor…
6333            67  ADDIS ABABA Ethiopia President Obama convene m…
6334            61  Jeb Bush suddenly attack Trump matter \n\n Jeb…

[6011 rows x 6 columns]
```

Splitting the dataset

[58]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
    df.lemmit_text, df.label,
    stratify=df.label, test_size=.2,
    random_state=42
)
```

[59]:
```
print(f'train shape:{X_train.shape}, test shape:{X_test.shape}')
```

```
train shape:(4808,), test shape:(1203,)
```

[60]:
```
y_train.value_counts()
```

[60]:
```
1    2420
0    2388
Name: label, dtype: int64
```

[61]:
```
y_test.value_counts()
```

[61]:
```
1    606
0    597
Name: label, dtype: int64
```

The last time it was used the CountVectorizer, now it will be tried the TF-IDF one.

[62]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

vectorizer = TfidfVectorizer().fit(X_train)

X_train = vectorizer.transform(X_train)
X_test = vectorizer.transform(X_test)
```

Which are the TF-IDF scores in the training set?

[63]:
```
vectorizer.vocabulary_
```

[63]:
```
{'seattle': 41025,
 'police': 35690,
 'black': 6566,
 'clothe': 9891,
 'day': 12415,
 'marcher': 28698,
 'hurl': 22480,
 'wrench': 50799,
 'rock': 39454,
 'officer': 33009,
 'hit': 21819,
 'stick': 43834,
 'friday': 18542,
 'night': 32126,
 'march': 28696,
 'neighborhood': 31789,
 'turn': 47169,
 'violent': 49264,
 'injure': 23591,
 'long': 27731,
 'demonstration': 12979,
 'management': 28533,
 'riot': 39288,
 'captain': 8390,
 'chris': 9441,
 'fowler': 18317,
 'tweet': 47221,
 'respond': 38856,
 'pepper': 34724,
 'spray': 43348,
 'ball': 5358,
 'eventually': 16406,
 'arrest': 4308,
 '15': 307,
 'people': 34717,
 'protester': 36778,
 'damage': 12204,
 'dozen': 14521,
 'vehicle': 48935,
 'say': 40531,
 'destructive': 13279,
 'chief': 9309,
 'kathleen': 25512,
 'toole': 46342,
 'joint': 24983,
```

```
'press': 36310,
'conference': 10614,
'mayor': 29135,
'ed': 15105,
'murray': 31204,
'think': 45857,
'professional': 36553,
'handle': 20830,
'situation': 42283,
'support': 44544,
'worker': 50698,
'right': 39233,
'cause': 8730,
'peaceful': 34575,
'include': 23171,
'lives': 27556,
'matter': 29047,
'immigrant': 22894,
'event': 16401,
'organize': 33402,
'group': 20270,
'el': 15306,
'comite': 10295,
'department': 13077,
'comprehensive': 10482,
'review': 39049,
'sure': 44578,
'use': 48595,
'force': 18134,
'appropriate': 4024,
'detail': 13290,
'injury': 23595,
'burn': 7909,
'orthopedic': 33469,
'conscious': 10758,
'good': 19808,
'spirit': 43255,
'bicycle': 6340,
'shadow': 41477,
'change': 9038,
'direction': 13676,
'capitol': 8371,
'hill': 21702,
'keep': 25598,
'interstate': 23959,
'away': 5013,
'downtown': 14513,
```

'gear': 19111,
'hem': 21444,
'plaza': 35475,
'central': 8877,
'college': 10158,
'reason': 37923,
'unify': 47971,
'jessica': 24802,
'ramirez': 37599,
'thousand': 45924,
'university': 48051,
'washington': 49777,
'student': 44129,
'diana': 13466,
'betancourt': 6233,
'take': 45035,
'deep': 12669,
'meaning': 29338,
'age': 2710,
'mom': 30576,
'sister': 42261,
'come': 10269,
'cross': 11737,
'arizona': 4220,
'desert': 13208,
'money': 30626,
'food': 18091,
'search': 41000,
'well': 50029,
'life': 27301,
'future': 18795,
'second': 41041,
'year': 51072,
'tuition': 47091,
'pay': 34530,
'make': 28410,
'dream': 14591,
'true': 46939,
'vote': 49458,
'technically': 45358,
'legal': 26980,
'alien': 3086,
'shameful': 41545,
'thing': 45854,
'grow': 20283,
'able': 2038,
'tell': 45449,

```
'identity': 22698,
'inside': 23678,
'secret': 41047,
'know': 26106,
'trust': 47009,
'go': 19679,
'help': 21429,
'report': 38673,
'crowd': 11776,
'federal': 17298,
'courthouse': 11409,
'stage': 43495,
'discuss': 13816,
'speaker': 43124,
'focus': 18030,
'deportation': 13111,
'break': 7374,
'family': 17047,
'everybody': 16422,
'proud': 36798,
'get': 19322,
'fight': 17559,
'jorge': 25019,
'baron': 5572,
'northwest': 32450,
'rights': 39245,
'project': 36616,
'win': 50412,
'justice': 25228,
'discussion': 13817,
'baltimore': 5389,
'state': 43630,
'attorney': 4765,
'marilyn': 28758,
'mosby': 30841,
'file': 17582,
'charge': 9086,
'early': 14970,
'morning': 30800,
'freddie': 18448,
'gray': 20083,
'case': 8595,
'indictment': 23307,
'cop': 11092,
'equity': 16040,
'oppress': 33303,
'nikita': 32153,
```

```
'oliver': 33118,
'demonstrator': 12982,
'humble': 22413,
'youth': 51205,
'street': 44043,
'nation': 31587,
'folk': 18058,
'general': 19148,
'view': 49183,
'27': 873,
'2016': 697,
'show': 41902,
'remain': 38519,
'arch': 4114,
'triumph': 46852,
'monument': 30724,
'destroy': 13273,
'islamic': 24302,
'militant': 30004,
'october': 32944,
'2015': 695,
'ancient': 3526,
'syrian': 44908,
'city': 9634,
'palmyra': 34068,
'video': 49163,
'luke': 27986,
'rudkowski': 39832,
'interview': 23971,
'matt': 29042,
'computer': 10502,
'programmer': 36599,
'allow': 3172,
'run': 39886,
'election': 15341,
'poll': 35745,
'800': 1651,
'000': 1,
'independently': 23279,
'finding': 17641,
'contradict': 10966,
'main': 28367,
'stream': 44035,
'medium': 29433,
'https': 22328,
'www': 50883,
'callforamerica': 8182,
```

```
'com': 10245,
'post': 35942,
'rigged': 39231,
'appear': 3955,
'raul': 37775,
'reyes': 39094,
'member': 29537,
'usa': 48567,
'today': 46243,
'board': 6849,
'contributor': 10991,
'follow': 18066,
'twitter': 47245,
'raulareyes': 37776,
'opinion': 33271,
'express': 16725,
'commentary': 10321,
'solely': 42844,
'author': 4878,
'cnn': 9957,
'prepare': 36239,
'july': 25173,
'fourth': 18314,
'weekend': 49961,
'grim': 20200,
'news': 31972,
'san': 40281,
'francisco': 18388,
'wednesday': 49950,
'kate': 25503,
'steinle': 43757,
'31': 972,
'fatally': 17168,
'shoot': 41848,
'apparently': 3949,
'randomly': 37638,
'walk': 49632,
'father': 17173,
'busy': 7974,
'pier': 35205,
'mexican': 29803,
'country': 11384,
'documentation': 14239,
'death': 12502,
'illegally': 22803,
'enter': 15886,
'deport': 13110,
```

```
'juan': 25105,
'lopez': 27785,
'sanchez': 40287,
'felony': 17381,
'accuse': 2247,
'horrific': 22173,
'crime': 11639,
'accord': 2215,
'immigration': 22897,
'authority': 4887,
'seven': 41402,
'conviction': 11041,
'drug': 14667,
'offense': 32995,
'symbol': 44845,
'estimate': 16247,
'11': 178,
'million': 30054,
'undocumented': 47801,
'united': 48029,
'states': 43642,
'poster': 35950,
'boy': 7227,
'control': 10999,
'illegal': 22801,
'southern': 43040,
'border': 7087,
'mexico': 29807,
'40': 1131,
'low': 27877,
'represent': 38690,
'overwhelming': 33852,
'majority': 28404,
'productive': 36540,
'society': 42784,
'simply': 42170,
'dangerous': 12248,
'individual': 23336,
'free': 18460,
'myth': 31349,
'increase': 23226,
'lead': 26865,
'research': 38767,
'policy': 35702,
'center': 8866,
'rate': 37740,
'fall': 17000,
```

```
'size': 42304,
'population': 35860,
'1990': 631,
'2010': 688,
'work': 50693,
'provide': 36815,
'consider': 10790,
'mass': 28951,
'shooting': 41850,
'aurora': 4851,
'newtown': 32017,
'charleston': 9106,
'commit': 10342,
'young': 51188,
'white': 50203,
'man': 28527,
'mean': 29333,
'potential': 35991,
'murderer': 31188,
'course': 11403,
'outlet': 33616,
'trumpet': 46962,
'murder': 31187,
'proof': 36669,
'criminal': 11646,
'overlook': 33754,
'ignore': 22760,
'story': 43958,
'genuine': 19211,
'hero': 21535,
'2013': 693,
'rescue': 38763,
'mother': 30875,
'child': 9316,
'staten': 43640,
'island': 24318,
'new': 31949,
'york': 51176,
'amidst': 3395,
'storm': 43952,
'surge': 44590,
'superstorm': 44525,
'sandy': 40330,
'takeaway': 45037,
'episode': 15997,
'possible': 35939,
'answer': 3719,
```

'problem': 36498,
'time': 46099,
'authorization': 4889,
'2011': 689,
'deputy': 13145,
'director': 13684,
'customs': 12040,
'enforcement': 15779,
'subcommittee': 44198,
'congress': 10697,
'cost': 11283,
'12': 224,
'500': 1278,
'person': 34870,
'multiply': 31133,
'taxpayer': 45282,
'waste': 49790,
'large': 26666,
'innocent': 23617,
'woman': 50609,
'lesson': 27132,
'need': 31742,
'smart': 42567,
'manpower': 28628,
'chase': 9132,
'productively': 36541,
'community': 10385,
'gardener': 18984,
'maid': 28347,
'felon': 17379,
'like': 27355,
'slip': 42502,
'crack': 11492,
'week': 49958,
'homeland': 22006,
'security': 41077,
'announce': 3673,
'rethink': 38940,
'priority': 36450,
'recent': 37996,
'arrival': 4314,
'step': 43775,
'start': 43605,
'seriously': 41341,
'target': 45184,
'real': 37870,
'threat': 45934,

'public': 36928,
'safety': 40095,
'government': 19912,
'effort': 15198,
'category': 8683,
'convict': 11039,
'crosser': 11746,
'terrorism': 45589,
'bar': 5480,
'err': 16118,
'seek': 41101,
'warrant': 49748,
'court': 11406,
'order': 33371,
'release': 38472,
'accordance': 2216,
'law': 26796,
'president': 36300,
'barack': 5482,
'obama': 32771,
'propose': 36705,
'executive': 16572,
'action': 2326,
'currently': 11999,
'tie': 46049,
'battle': 5726,
'difference': 13537,
'resource': 38837,
'tragedy': 46545,
'allege': 3128,
'deserve': 13214,
'vilify': 49214,
'false': 17021,
'association': 4575,
'look': 27752,
'forward': 18277,
'bubble': 7718,
'burst': 7926,
'bay': 5752,
'area': 4160,
'little': 27532,
'high': 21670,
'tech': 45352,
'weenie': 49969,
'bug': 7773,
'techie': 45353,
'rumor': 39879,

```
'round': 39697,
'layoff': 26835,
'grant': 20024,
'befuddle': 5922,
'tax': 45271,
'employment': 15643,
'taxis': 45278,
'inopportune': 23633,
'moment': 30579,
'glory': 19624,
'commercial': 10327,
'estate': 16234,
'workforce': 50702,
'300': 950,
'bloomberg': 6764,
'throw': 45969,
'183': 448,
'642': 1453,
'square': 43402,
'foot': 18108,
'vacant': 48711,
'office': 33007,
'space': 43064,
'build': 7782,
'mid': 29914,
'market': 28796,
'headquarters': 21221,
'sublease': 44224,
'bring': 7512,
'51': 1300,
'msf': 31003,
'this': 45880,
'snapshot': 42663,
'cushman': 12025,
'wakefield': 49605,
'leasing': 26911,
'activity': 2339,
'nearly': 31720,
'grind': 20208,
'halt': 20743,
'quarter': 37274,
'875': 1719,
'sf': 41445,
'lease': 26908,
'2001': 675,
'there': 45789,
'major': 28400,
```

```
'deal': 12480,
'100': 124,
'amazon': 3324,
'live': 27545,
'platform': 35441,
'twitch': 47242,
'178': 403,
'half': 20708,
'wework': 50111,
'78': 1608,
'plunge': 35554,
'30': 949,
'period': 34788,
'service': 41355,
'firm': 17713,
'savills': 40504,
'studley': 44137,
'add': 2388,
'dryly': 14690,
'competition': 10429,
'calm': 8190,
'dramatically': 14564,
'ago': 2761,
'and': 3529,
'lot': 27821,
'supply': 44543,
'construction': 10853,
'prelease': 36206,
'overall': 33696,
'vacancy': 48710,
'rise': 39306,
'percentage': 34735,
'point': 35644,
'prior': 36443,
'q3': 37151,
'class': 9719,
'building': 7785,
'availability': 4944,
'jump': 25178,
'10': 123,
'red': 38115,
'hot': 22220,
'cold': 10110,
'soma': 42889,
'practically': 36061,
'end': 15716,
'spectrum': 43176,
```

'financial': 17632,
'district': 14081,
'south': 43033,
'spike': 43236,
'way': 49855,
'link': 27438,
'wolfstreet': 50603,
'alive': 3110,
'lawn': 26812,
'care': 8444,
'maintenance': 28381,
'previous': 36368,
'page': 33971,
'bamacare': 5399,
'supreme': 44565,
'gut': 20494,
'health': 21235,
'leave': 26917,
'americans': 3373,
'face': 16865,
'severe': 41409,
'consequence': 10770,
'king': 25929,
'burwell': 7932,
'lawsuit': 26818,
'originate': 33431,
'conservative': 10779,
'libertarian': 27251,
'tank': 45137,
'stray': 44030,
'phrase': 35142,
'affordable': 2648,
'act': 2322,
'exchange': 16533,
'establish': 16225,
'subsidy': 44265,
'resident': 38798,
'refuse': 38295,
'insurance': 23776,
'13': 263,
'columbia': 10235,
'bid': 6341,
'derail': 13150,
'succeed': 44301,
'disappear': 13717,
'maybe': 29125,
'immediately': 22887,

```
'later': 26722,
'obamacare': 32774,
'enrollee': 15862,
'number': 32644,
'human': 22389,
'moderate': 30468,
'income': 23185,
'recipient': 38017,
'result': 38902,
'close': 9878,
'lose': 27814,
'coverage': 11437,
'premium': 36222,
'huffington': 22363,
'risk': 39312,
'bad': 5223,
'effect': 15185,
'ruling': 39864,
'want': 49698,
'affect': 2623,
'absence': 2097,
'ruin': 39853,
'lifelong': 27309,
'plan': 35408,
'jeopardy': 24781,
'disrupt': 14004,
'tuesday': 47085,
'amazing': 3322,
'hillary': 21704,
'clinton': 9837,
'florida': 17947,
'ohio': 33057,
'bernie': 6180,
'sanders': 40312,
'path': 34419,
'nomination': 32295,
'impossible': 23019,
'marco': 28705,
'rubio': 39808,
'drop': 14649,
'race': 37420,
'virtually': 49285,
'ensure': 15877,
'donald': 14335,
'trump': 46953,
'ted': 45376,
'cruz': 11840,
```

'couple': 11393,
'month': 30711,
'strength': 44050,
'conventional': 11017,
'wisdom': 50496,
'grab': 19940,
'republican': 38724,
'voter': 49462,
'lapel': 26644,
'scream': 40894,
'idiot': 22712,
'realize': 37887,
'democrats': 12951,
'scared': 40608,
'guy': 20513,
'write': 50821,
'yglesias': 51125,
'january': 24617,
'see': 41094,
'nimble': 32168,
'politician': 35723,
'reassure': 37939,
'appeal': 3952,
'latinos': 26733,
'blunt': 6815,
'turnout': 47179,
'fractious': 18355,
'party': 34346,
'poor': 35839,
'performance': 34773,
'primary': 36406,
'split': 43290,
'weak': 49884,
'candidate': 8284,
'delusion': 12891,
'behalf': 5945,
'political': 35719,
'establishment': 16228,
'pretend': 36338,
'bend': 6071,
'strong': 44099,
'far': 17080,
'unacceptable': 47421,
'base': 5634,
'hold': 21941,
'season': 41012,
'weakness': 49890,

```
'hopeful': 22125,
'angry': 3618,
'record': 38076,
'occasional': 32905,
'pivot': 35367,
'particularly': 34321,
'résumé': 39990,
'relatively': 38461,
'thin': 45852,
'remind': 38548,
'republicans': 38726,
'criticism': 11696,
'attack': 4733,
'acceptable': 2177,
'wing': 50441,
'choice': 9396,
'gop': 19839,
'middle': 29921,
'rhetoric': 39120,
'partisan': 34330,
'fearmongere': 17273,
'crucially': 11794,
'speed': 43193,
'speak': 43123,
'language': 26621,
'optimism': 33325,
'uplift': 48468,
'talk': 45074,
'economy': 15087,
'argument': 4185,
'divisive': 14138,
'critique': 11700,
'past': 34388,
'fit': 17750,
'mood': 30730,
'skill': 42351,
'traditionally': 46535,
'kind': 25914,
'hardcore': 20925,
'ideologue': 22704,
'nominate': 32292,
'barry': 5605,
'goldwater': 19784,
'extremist': 16794,
'extraordinarily': 16778,
'unfavorable': 47871,
'inspire': 23713,
```

'hispanics': 21802,
'solve': 42881,
'enthusiasm': 15902,
'gap': 18966,
'compare': 10405,
'campaign': 8234,
'dare': 12297,
'hope': 22124,
'suit': 44371,
'2000': 672,
'george': 19246,
'bush': 7941,
'seemingly': 41106,
'governor': 19915,
'texas': 45644,
'compassionate': 10414,
'2008': 683,
'john': 24956,
'mccain': 29177,
'independent': 23278,
'kerry': 25692,
'try': 47037,
'democratic': 12943,
'ticket': 46039,
'2012': 690,
'mitt': 30390,
'romney': 39556,
'blue': 6789,
'predict': 36155,
'american': 3366,
'politic': 35718,
'magical': 28281,
'land': 26576,
'surprise': 44606,
'ask': 4473,
'opponent': 33287,
'outcome': 33586,
'give': 19507,
'chance': 9029,
'senate': 41224,
'pull': 36976,
'landslide': 26605,
'rarely': 37702,
'modern': 30474,
'spotlight': 43334,
'global': 19595,
'warming': 49730,

```
'saturday': 40459,
'travel': 46686,
'alaska': 2971,
'alaskans': 2973,
'weekly': 49963,
'address': 2407,
'experience': 16665,
'wildfire': 50350,
'expect': 16639,
'average': 4966,
'temperature': 45470,
'degree': 12794,
'climate': 9813,
'village': 49218,
'imminent': 22898,
'danger': 12247,
'sea': 40968,
'water': 49815,
'ice': 22651,
'glacier': 19529,
'melt': 29529,
'happen': 20891,
'fellow': 17376,
'threaten': 45935,
'wipe': 50480,
'town': 46470,
'power': 36026,
'protect': 36755,
'pose': 35917,
'strike': 44075,
'tricky': 46808,
'balance': 5331,
'environmental': 15947,
'conservation': 10777,
'energy': 15769,
'production': 36538,
'expand': 16628,
'oil': 33065,
'drilling': 14623,
'alaskan': 2972,
'coast': 9985,
'fuel': 18678,
'contribute': 10987,
'sharp': 41610,
'environmentalist': 15949,
'trip': 46836,
'begin': 5933,
```

```
'monday': 30608,
'activist': 2335,
'organization': 33398,
'credo': 11591,
'visit': 49309,
'self': 41165,
'defeat': 12695,
'hypocrisy': 22592,
'leader': 26868,
'drill': 14622,
'arctic': 4152,
'online': 33199,
'petition': 34961,
'urgency': 48538,
'massive': 28963,
'fossil': 18283,
'extraction': 16766,
'leadership': 26873,
'share': 41587,
'concern': 10537,
'offshore': 33028,
'note': 32488,
'remember': 38541,
'bp': 7247,
'spill': 43237,
'gulf': 20431,
'rely': 38516,
'gas': 19024,
'transition': 46624,
'renewable': 38597,
'source': 43023,
'wind': 50415,
'solar': 42837,
'believe': 6011,
'domestic': 14312,
'foreign': 18165,
'import': 23007,
'demand': 12904,
'standard': 43546,
'industry': 23378,
'administration': 2472,
'issue': 24365,
'permit': 34823,
'shell': 41677,
'mandate': 28553,
'strict': 44066,
'company': 10398,
```

```
'meet': 29448,
'testament': 45612,
'rigorous': 39252,
'apply': 3995,
'delay': 12825,
'limit': 27395,
'exploration': 16702,
'line': 27418,
'continue': 10938,
'america': 3362,
'precious': 36127,
'schedule': 40663,
'participate': 34311,
'anchorage': 3524,
'tour': 46450,
'coastal': 9986,
'fishing': 17734,
'globe': 19610,
'wakeup': 49608,
'world': 50718,
'late': 26714,
'colorado': 10209,
'convention': 11016,
'springs': 43361,
'supporter': 44547,
'wave': 49846,
'big': 6359,
'broom': 7625,
'letter': 27145,
'fasten': 17157,
'place': 35387,
'hockey': 21888,
'arena': 4164,
'prop': 36673,
'probably': 36489,
'familiar': 17042,
'sport': 43326,
'fan': 17051,
'sweep': 44754,
'pick': 35177,
'34': 1022,
'national': 31588,
'delegate': 12829,
'award': 5008,
'back': 5170,
'spot': 43333,
'congressional': 10700,
```

'statewide': 43651,
'slot': 42524,
'similar': 42142,
'north': 32438,
'dakota': 12180,
'elect': 15335,
'approve': 4031,
'slate': 42439,
'haul': 21107,
'complex': 10451,
'maneuvering': 28568,
'louisiana': 27847,
'caucus': 8718,
'highlight': 21676,
'organizational': 33399,
'frontrunner': 18619,
'recently': 37997,
'insist': 23695,
'237': 815,
'clinch': 9824,
'outright': 33646,
'denver': 13068,
'contested': 10925,
'significant': 42089,
'possibility': 35938,
'confidence': 10627,
'scenario': 40631,
'senator': 41225,
'stop': 43929,
'success': 44303,
'offer': 33000,
'block': 6718,
'short': 41866,
'eventual': 16404,
'nominee': 32296,
'observer': 32864,
'figure': 17571,
'play': 35457,
'role': 39517,
'decline': 12620,
'bind': 6432,
'congressman': 10702,
'ken': 25639,
'buck': 7727,
'chair': 8983,
'volunteer': 49438,
'december': 12575,

```
 'vet': 49083,
 'local': 27628,
 'regional': 38327,
 'meeting': 29449,
 'wear': 49904,
 'bright': 7497,
 'orange': 33345,
 'shirt': 41803,
 'hand': 20805,
 'glossy': 19628,
 'sheet': 41659,
 'list': 27487,
 'preferred': 36194,
 'blast': 6648,
 'text': 45647,
 …}
```

[64]:
```
names = vectorizer.get_feature_names_out()
scores_dict = {}
words = []
scores = []

for word in names:
    key = vectorizer.vocabulary_.get(word)
    score = vectorizer.idf_[key]
    words.append(word)
    scores.append(score)

scores_dict['tf_idf_word'] = words
scores_dict['tf_idf_score'] = scores
```

[65]:
```
sorted_scores = pd.DataFrame(scores_dict).sort_values('tf_idf_score',␣
 ↪ascending=False)
sorted_scores
```

[65]:

	tf_idf_word	tf_idf_score
25983	kiryat	8.785097
35209	pierini	8.785097
35213	pies	8.785097
35214	piet	8.785097
17432	ferryman	8.785097
…	…	…
31949	new	1.606170
51072	year	1.573172
34717	people	1.572434
46099	time	1.567654
40531	say	1.349364

```
[51967 rows x 2 columns]
```

5 Model Creation

It will be used the MultinomialNB.

[66]:
```
from sklearn.naive_bayes import MultinomialNB
```

[67]:
```
model = MultinomialNB()

model.fit(X_train, y_train)
```

[67]:
```
MultinomialNB()
```

5.1 Results

It will be printed a classification report, but even other metrics can be computed (e.g. ROC AUC).

[68]:
```
from sklearn.metrics import classification_report

print(classification_report(y_test, model.predict(X_test)))
```

```
              precision    recall  f1-score   support

           0       0.73      0.98      0.84       597
           1       0.97      0.64      0.77       606

    accuracy                           0.81      1203
   macro avg       0.85      0.81      0.80      1203
weighted avg       0.85      0.81      0.80      1203
```

Results seem worse from the first attempt.

Other approaches could be:

trying tf-idf on the original text
trying CountVectorizer on the preprocessed text
trying Spacy word vectors

6 Trying TF-IDF without Lemmatization

[69]:
```
df
```

[69]:
```
                                                title  \
0                           You Can Smell Hillary's Fear
1       Watch The Exact Moment Paul Ryan Committed Pol…
2             Kerry to go to Paris in gesture of sympathy
```

```
3     Bernie supporters on Twitter erupt in anger ag…
4      The Battle of New York: Why This Primary Matters
…                                                    …
6330  State Department says it can't find emails fro…
6331  The ‘P’ in PBS Should Stand for ‘Plutocratic’ …
6332  Anti-Trump Protesters Are Tools of the Oligarc…
6333  In Ethiopia, Obama seeks progress on peace, se…
6334  Jeb Bush Is Suddenly Attacking Trump. Here's W…

                                                 text  label  text_length  \
0     Daniel Greenfield, a Shillman Journalism Fello…      1         7518
1     Google Pinterest Digg Linkedin Reddit Stumbleu…      1         2646
2     U.S. Secretary of State John F. Kerry said Mon…      0         2543
3     - Kaydee King (@KaydeeKing) November 9, 2016 T…      1         2660
4     It's primary day in New York and front-runners…      0         1840
…                                                    …    …           …
6330  The State Department told the Republican Natio…      0         4076
6331  The ‘P’ in PBS Should Stand for ‘Plutocratic’ …      1        14323
6332   Anti-Trump Protesters Are Tools of the Oligar…      1        11974
6333  ADDIS ABABA, Ethiopia -President Obama convene…      0         6991
6334  Jeb Bush Is Suddenly Attacking Trump. Here's W…      0         4818

      title_length                                       lemmit_text
0               28  Daniel Greenfield Shillman Journalism Fellow F…
1               85  Google Pinterest Digg Linkedin Reddit Stumbleu…
2               43  U.S. Secretary State John F. Kerry say Monday …
3               84  Kaydee King @KaydeeKing November 9 2016 lesson…
4               48  primary day New York runner Hillary Clinton Do…
…              …                                                 …
6330            69  State Department tell Republican National Comm…
6331            59  p PBS stand plutocratic Pentagon post Oct 27 2…
6332            66    Anti Trump Protesters Tools Oligarchy  refor…
6333            67  ADDIS ABABA Ethiopia President Obama convene m…
6334            61  Jeb Bush suddenly attack Trump matter \n\n Jeb…

[6011 rows x 6 columns]
```

[77]:
```
X_train, X_test, y_train, y_test = train_test_split(
    df.text, df.label,
    stratify = df.label, random_state=42
)
```

[78]:
```
y_train.value_counts()
```

[78]:
```
1    2269
0    2239
Name: label, dtype: int64
```

[79]:
```
y_test.value_counts()
```

[79]:
```
1    757
0    746
Name: label, dtype: int64
```

Vectorizing

[80]:
```
vectorizer = TfidfVectorizer().fit(X_train)

X_train = vectorizer.transform(X_train)
X_test = vectorizer.transform(X_test)
```

Model creation

[81]:
```
model = MultinomialNB()

model.fit(X_train, y_train)
```

[81]:
```
MultinomialNB()
```

Results

[82]:
```
print(classification_report(y_test, model.predict(X_test)))
```

```
              precision    recall  f1-score   support

           0       0.69      0.99      0.81       746
           1       0.98      0.56      0.71       757

    accuracy                           0.77      1503
   macro avg       0.84      0.77      0.76      1503
weighted avg       0.84      0.77      0.76      1503
```

Without feature engineering results have worsened a lot.

7 Trying CountVectorizer on the Preprocessed Text

[83]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
    df.lemmit_text, df.label,
    stratify=df.label, test_size=.2,
    random_state=42
)
```

[85]:
```
y_train.value_counts()
```

[85]:
```
1    2420
0    2388
Name: label, dtype: int64
```

[86]:
```
y_test.value_counts()
```

[86]:
```
1    606
0    597
Name: label, dtype: int64
```

Vectorizing

[87]:
```
from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer().fit(X_train)

X_train = vectorizer.transform(X_train)
X_test = vectorizer.transform(X_test)
```

Model

[88]:
```
model = MultinomialNB()

model.fit(X_train, y_train)
```

[88]:
```
MultinomialNB()
```

Results

[89]:
```
print(classification_report(y_test, model.predict(X_test)))
```

```
              precision    recall  f1-score   support

           0       0.85      0.93      0.89       597
           1       0.93      0.84      0.88       606

    accuracy                           0.88      1203
   macro avg       0.89      0.88      0.88      1203
weighted avg       0.89      0.88      0.88      1203
```

Results are similar to the first attempt without preprocessing text.

8 Using Spacy without Preprocessing

[2]:
```
nlp = spacy.load('en_core_web_lg')
```

Here VSCode crashed, it is always important to save files after preprocessing.

[3]:
```
df = pd.read_csv('preprocessed_fake_news.csv')
df
```

[3]:
```
                                                  title  \
0                          You Can Smell Hillary's Fear
1     Watch The Exact Moment Paul Ryan Committed Pol…
2           Kerry to go to Paris in gesture of sympathy
3     Bernie supporters on Twitter erupt in anger ag…
4      The Battle of New York: Why This Primary Matters
…                                                   …
6006  State Department says it can't find emails fro…
6007  The 'P' in PBS Should Stand for 'Plutocratic' …
6008  Anti-Trump Protesters Are Tools of the Oligarc…
6009  In Ethiopia, Obama seeks progress on peace, se…
6010  Jeb Bush Is Suddenly Attacking Trump. Here's W…

                                                  text  label  text_length  \
0     Daniel Greenfield, a Shillman Journalism Fello…      1         7518
1     Google Pinterest Digg Linkedin Reddit Stumbleu…      1         2646
2     U.S. Secretary of State John F. Kerry said Mon…      0         2543
3     - Kaydee King (@KaydeeKing) November 9, 2016 T…      1         2660
4     It's primary day in New York and front-runners…      0         1840
…                                                   …    …           …
6006  The State Department told the Republican Natio…      0         4076
6007  The 'P' in PBS Should Stand for 'Plutocratic' …      1        14323
6008   Anti-Trump Protesters Are Tools of the Oligar…      1        11974
6009  ADDIS ABABA, Ethiopia -President Obama convene…      0         6991
6010  Jeb Bush Is Suddenly Attacking Trump. Here's W…      0         4818

      title_length                                       lemmit_text
0               28  Daniel Greenfield Shillman Journalism Fellow F…
1               85  Google Pinterest Digg Linkedin Reddit Stumbleu…
2               43  U.S. Secretary State John F. Kerry say Monday …
3               84  Kaydee King @KaydeeKing November 9 2016 lesson…
4               48  primary day New York runner Hillary Clinton Do…
…              …                                                 …
6006            69  State Department tell Republican National Comm…
6007            59  p PBS stand plutocratic Pentagon post Oct 27 2…
6008            66    Anti Trump Protesters Tools Oligarchy  refor…
6009            67  ADDIS ABABA Ethiopia President Obama convene m…
6010            61  Jeb Bush suddenly attack Trump matter \n\n Jeb…

[6011 rows x 6 columns]
```

Creating a Spacy vector.

[4]:
```
df['spacy_vector'] = df.text.apply(lambda x: nlp(x).vector)
df
```

[4]:

	title	text	label	text_length	title_length	lemmit_text	spacy_vector
0	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello…	1	7518	28	Daniel Greenfield Shillman Journalism Fellow F…	[-1.3751823, 1.3421849, -2.3666484, 0.12908486…
1	Watch The Exact Moment Paul Ryan Committed Pol…	Google Pinterest Digg Linkedin Reddit Stumbleu…	1	2646	85	Google Pinterest Digg Linkedin Reddit Stumbleu…	[-1.7449774, 0.93961924, -2.024867, 0.42536643…
2	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon…	0	2543	43	U.S. Secretary State John F. Kerry say Monday …	[-1.9426425, 1.0062195, -1.9992222, 0.20469022…
3	Bernie supporters on Twitter erupt in anger ag…	- Kaydee King (@KaydeeKing) November 9, 2016 T…	1	2660	84	Kaydee King @KaydeeKing November 9 2016 lesson…	[-1.9125352, -0.1481846, -1.1432766, 0.6861217…
4	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners…	0	1840	48	primary day New York runner Hillary Clinton Do…	[-1.8516092, 1.3163909, -2.1726575, 1.2286776,…
…	…	…	…	…	…	…	…
6006	State Department says it can't find emails fro…	The State Department told the Republican Natio…	0	4076	69	State Department tell Republican National Comm…	[-1.556691, 0.60453945, -1.1016529, 0.16134764…
6007	The 'P' in PBS Should Stand for 'Plutocratic' …	The 'P' in PBS Should Stand for 'Plutocratic' …	1	14323	59	p PBS stand plutocratic Pentagon post Oct 27 2…	
6008	Anti-Trump Protesters Are Tools of the Oligarc…	Anti-Trump Protesters Are Tools of the Oligar…	1	11974	66	Anti Trump Protesters Tools Oligarchy refor…	
6009	In Ethiopia, Obama seeks progress on peace, se…	ADDIS ABABA, Ethiopia -President Obama convene…	0	6991	67	ADDIS ABABA Ethiopia President Obama convene m…	
6010	Jeb Bush Is Suddenly Attacking Trump. Here's W…	Jeb Bush Is Suddenly Attacking Trump. Here's W…	0	4818	61	Jeb Bush suddenly attack Trump matter \n\n Jeb…	

```
6007  [-2.206026, -0.12085343, -1.0834901, 0.6655213…
6008  [-2.558132, 0.47698027, -1.8662019, 0.8344748,…
6009  [-1.8501893, 0.6893597, -1.9594386, 0.41594952…
6010  [-1.4473745, 2.095408, -2.1326072, 0.4212142, …

[6011 rows x 7 columns]
```

Saving the model with the new feature.

[5]:
```
df.to_csv('preprocessed_fake_news.csv', index=None)
```

[6]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
    df.spacy_vector, df.label,
    stratify=df.label, random_state=42
)
```

[9]:
```
X_train
```

[9]:
```
4890    [-2.0016534, 0.007585724, -0.8894147, 0.046493…
4682    [-2.5661526, 0.43509185, -1.8436817, 0.8052051…
821     [-0.20114365, -0.7524439, -1.3927656, -1.92348…
38      [-1.2550871, 0.8170438, -1.2368673, 0.07717796…
1637    [-1.7568897, -0.12037473, -0.8760516, 0.601860…
                                …
4886    [-1.8356636, 0.4669065, -1.7419188, 0.19650374…
4288    [-1.2327241, 0.8559157, -2.7709413, -0.4093651…
257     [-2.1132298, -0.17463753, -0.46453714, 0.28716…
5294    [-2.188368, 0.72217894, -1.6713092, 0.6883849,…
1849    [-1.7973229, 0.2899992, -1.4270326, 0.16144261…
Name: spacy_vector, Length: 4508, dtype: object
```

[10]:
```
y_train.value_counts()
```

[10]:
```
1    2269
0    2239
Name: label, dtype: int64
```

[11]:
```
y_test.value_counts()
```

[11]:
```
1    757
0    746
Name: label, dtype: int64
```

In order to work properly, the X_train should be transformed.

[14]:
```
X_train = np.stack(X_train)
X_test = np.stack(X_test)
```

Moreover, MultinomialNB does not accept negative values.
They will be scaled using MinMaxScaler

[16]:
```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler().fit(X_train)

X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

Model

[21]:
```
from sklearn.naive_bayes import MultinomialNB

model = MultinomialNB()

model.fit(X_train, y_train)
```

[21]:
```
MultinomialNB()
```

Results

[23]:
```
from sklearn.metrics import classification_report

print(classification_report(y_test, model.predict(X_test)))
```

```
              precision    recall  f1-score   support

           0       0.76      0.84      0.80       746
           1       0.82      0.74      0.78       757

    accuracy                           0.79      1503
   macro avg       0.79      0.79      0.79      1503
weighted avg       0.79      0.79      0.79      1503
```

9 Conclusion

I found very useful trying different ways of preprocessing to see how the model perform using different data.

A further implementation can be to try different classification models on the same data and providing more metrics.

Even the feature engineering could be performed using different NLP tools.