509 Final Project

The notebook is for Exploratory Data Analysis (EDA), text data preprocessing, modeling, and evaluation.

Globally import libraries

In [1]:
```
from bs4 import BeautifulSoup
from collections import defaultdict, Counter
import datetime as dt
import emoji
import itertools
import json
import logging
import matplotlib.pyplot as plt
import numpy as np
import os
import pandas as pd
import pickle
import pymysql as mysql
import random
import re
import regex as rex
import requests
import shutil
from string import punctuation
import time
from tqdm import tqdm
import zipfile

import nltk
from nltk.corpus import stopwords
import spacy

from skopt import BayesSearchCV
from skopt.space import Real, Categorical, Integer

from sklearn.feature_extraction.text import TfidfTransformer, \
CountVectorizer, TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.pipeline import make_pipeline, Pipeline
from sklearn import metrics
from sklearn.metrics import make_scorer, f1_score, classification_report, \
confusion_matrix, ConfusionMatrixDisplay, RocCurveDisplay

import textacy.preprocessing as tprep
from textacy.extract import keyword_in_context

# Set pandas global options
```

```
pd.options.display.max_rows = 17
pd.options.display.precision = 4
np.set_printoptions(suppress=True, precision=4)

%matplotlib inline
```

Upload data from CSV

In [2]:
```
'''Dir nav citation:
https://softhints.com/python-change-directory-parent/'''
curr_dir = os.path.abspath(os.curdir)
print(curr_dir)
os.chdir("..")
up1_dir = os.path.abspath(os.curdir)
print(up1_dir)
```

```
C:\Users\acarr\Documents\GitHub\ADS509_Final_project\deliverables
C:\Users\acarr\Documents\GitHub\ADS509_Final_project
```

In [3]:
```
# change `data_location` to the location of the folder on your machine.
data_location = 'data'

file_in_name01 = 'master.csv'
file_in_name02 = 'master_business_TheHill.csv'

file_in_path01 = os.path.join(up1_dir, data_location, file_in_name01)
file_in_path02 = os.path.join(up1_dir, data_location, file_in_name02)

print(f'CSV file 1 in path: {file_in_path01}')
print(f'CSV file 2 in path: {file_in_path02}')
```

```
CSV file 1 in path: C:\Users\acarr\Documents\GitHub\ADS509_Final_project\data\maste
r.csv
CSV file 2 in path: C:\Users\acarr\Documents\GitHub\ADS509_Final_project\data\master
_business_TheHill.csv
```

Review dataframe

In [4]:
```
slct_tbl_full_df01 = pd.read_csv(file_in_path01)
print(f'Dataframe shape: {slct_tbl_full_df01.shape}')
display(slct_tbl_full_df01.head())
```

```
Dataframe shape: (4509, 7)
```

	source_name	author	title	url	publis
0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/...	20 30T16:
1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0...	20 30T19:
2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/...	20 30T23:
3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0...	20 30T18:
4	The Washington Post	NaN	The revolt of Christian home-schoolers...	https://www.washingtonpost.com/education/inter...	20 30T18:

Exploratory Data Analysis (EDA)

Count missing `article_text` feature

The majority of null values appear in the `content` column. There are also several in `author` and one in `article_text`. Neither `content` nor `author` will be used for current modeling efforts, therefore they are not a factor. The one instance with missing article text will be removed.

In [5]:
```
count_nan = slct_tbl_full_df01.isnull().sum()

# printing the number of values present
# in the column
print('Number of NaN values present: ' + str(count_nan))
```

```
Number of NaN values present: source_name        0
author            37
title              0
url                0
publish_date       0
content         3351
article_text       1
dtype: int64
```

Count blank `article_text` feature

In [6]:
```
print(len(slct_tbl_full_df01[slct_tbl_full_df01['article_text']=='']))
display(slct_tbl_full_df01[slct_tbl_full_df01['article_text']=='].head(20))
```

0

source_name	author	title	url	publish_date	content	article_text

Remove missing `article_text` row(s)

In [7]:
```
'''Drop missing citation:
https://pandas.pydata.org/pandas-docs/stable/reference
/api/pandas.DataFrame.dropna.html#pandas.DataFrame.dropna'''
slct_tbl_full_df02 = slct_tbl_full_df01.dropna(subset=['article_text'])
print(f'Dataframe shape: {slct_tbl_full_df02.shape}')
display(slct_tbl_full_df02.head())
```

```
Dataframe shape: (4508, 7)
```

	source_name	author	title	url	pub
0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/...	30T
1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0...	30T
2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/...	30T
3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0...	30T
5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0...	30T

Count characters and words for initial review

In [8]:
```
tqdm.pandas(ncols=50)  # can use tqdm_gui, optional kwargs, etc
# Now you can use `progress_apply` instead of `apply`

# Raw text character and word counts
slct_tbl_full_df02['char_cnt'] = slct_tbl_full_df02['article_text']\
.progress_apply(len)
slct_tbl_full_df02['word_cnt'] = slct_tbl_full_df02['article_text']\
.progress_apply(lambda x: len(x.split()))
display(slct_tbl_full_df02.head())
```

```
100%|██████| 4508/4508 [00:00<00:00, 322023.34it/s]
C:\Users\acarr\AppData\Local\Temp\ipykernel_23812\2936833956.py:5: SettingWithCopyWa
rning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  slct_tbl_full_df02['char_cnt'] = slct_tbl_full_df02['article_text']\
100%|██████| 4508/4508 [00:00<00:00, 11534.89it/s]
C:\Users\acarr\AppData\Local\Temp\ipykernel_23812\2936833956.py:7: SettingWithCopyWa
rning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  slct_tbl_full_df02['word_cnt'] = slct_tbl_full_df02['article_text']\
```

	source_name	author	title	url	pub
0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/...	30T
1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0...	30T
2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/...	30T
3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0...	30T
5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0...	30T

Descriptive statistics

Stats are displayed for both categorical and numerical columns. As expected "Fox News" is the most frequent value in `source_name` as the most articles were collected from that news site. The inclusion of "Associated Press" as the mode for `author` identified it as a potential source for skew in the final results, as AP was rated as a "center" source in the AllSide Media Bias Chart. As a result, all articles with an `author` value of "Associated Press" were removed; similarily, articles by "msn" and "Reuters" were also removed.
For the numerical values, there was a very large range for both character and word counts (80,454 and 14,306, respectively), but also a large delta between the 75% percentile and max (74,920.5 and 13,433, respectively), indicating a distribution with a very long right tail with a very small amount of some extremely long (outlier) articles. As a result, the standard deviation was also quite large relative to the mean. For the current analyses, no additional efforts will be performed to account for outliers, but this will be an examination factor for future expansion/comparative studies.

In [9]:
```
slct_tbl_full_df02[['source_name',
                    'author',
                    'publish_date',
                    'article_text']].describe(include="O").T
```

Out[9]:

	count	unique	top	freq
source_name	4508	4	Fox News	2192
author	4472	956	Associated Press	450
publish_date	4508	4486	2023-05-13T11:00:00Z	3
article_text	4508	4508	Travelers in Alabama driving on Interstate 65 ...	1

In [10]:
```
slct_tbl_full_df02.describe().T
```

Out[10]:

	count	mean	std	min	25%	50%	75%	max
char_cnt	4508.0	4655.5011	3137.3650	131.0	2832.0	3951.5	5664.5	80585.0
word_cnt	4508.0	731.0315	518.5765	16.0	432.0	607.0	889.0	14322.0

Display Source counts

In [11]:
```
slct_tbl_full_df02['source_name'].value_counts()
```

Out[11]:
```
Fox News               2192
Breitbart News         1017
CNN                     773
The Washington Post     526
Name: source_name, dtype: int64
```

Examine inclusion of "centrist" sources indicated by `author` feature

In [12]:
```
slct_tbl_full_df02a = slct_tbl_full_df02[slct_tbl_full_df02['author']\
                                         .isin(['msn',
                                                'Associated Press',
                                                'Reuters'])]

display(slct_tbl_full_df02a[slct_tbl_full_df02a['author']=='msn'])

display(slct_tbl_full_df02a.groupby(by=['source_name', 'author']).count())
```

	source_name	author	title	url	pu
17	The Washington Post	msn	State Dept seeks to expand space diplomacy...	https://www.washingtonpost.com/technology/2023...	30
18	The Washington Post	msn	SHOCK IN RUSSIAN CAPITAL	https://www.washingtonpost.com/world/2023/05/3...	30
22	The Washington Post	msn	Debate over whether AI will destroy us is divi...	https://www.washingtonpost.com/technology/2023...	20
81	The Washington Post	msn	Corporate bankruptcies creeping up as pressure...	https://www.washingtonpost.com/business/2023/0...	23
84	The Washington Post	msn	The looming existential crisis for cable news...	https://www.washingtonpost.com/media/2023/05/2...	23
...	...	...	...	...	
492	The Washington Post	msn	Biden shows growing appetite to cross Putin's ...	https://www.washingtonpost.com/national-securi...	01
502	The Washington Post	msn	Behind-the-scenes videos of Tucker Carlson wer...	https://www.washingtonpost.com/media/2023/06/0...	02
503	The Washington Post	msn	Georgia probe of Trump broadens to activities ...	https://www.washingtonpost.com/nation/2023/06/...	02
506	The Washington Post	msn	DRAMA: Couple, both nurses, save man's life mi...	https://www.washingtonpost.com/lifestyle/2023/...	02

	source_name	author	title	url	pu
509	The Washington Post	msn	'DRAG RACE' queen says cancellation of militar...	https://www.washingtonpost.com/nation/2023/06/...	02

25 rows × 9 columns

		title	url	publish_date	content	article_text	char_cnt	word_cnt
source_name	**author**							
CNN	**Reuters**	6	6	6	1	6	6	6
Fox News	**Associated Press**	450	450	450	73	450	450	450
	Reuters	1	1	1	0	1	1	1
The Washington Post	**msn**	25	25	25	25	25	25	25

In [13]:
```
counter = Counter(slct_tbl_full_df02['author'])

word_cutoff = 5
con_feature_words = set()

for word, count in counter.items():
    if count > word_cutoff:
        con_feature_words.add(word)

print(f'''With a word cutoff of {word_cutoff}, we have
{len(con_feature_words)} words as features in the model.''')
print(con_feature_words)
```

```
With a word cutoff of 5, we have
151 words as features in the model.
{nan, 'Greg Gutfeld', 'Julia Musto', 'Ryan Morik', 'Devlin Barrett', 'Associated Pre
ss', 'Hanna Panreck', 'Howard Kurtz', 'Lawrence Richard', 'Amy B Wang', 'Gabriel Hay
s', 'Breitbart London, Breitbart London', 'Bradford Betz', 'Timothy Nerozzi', 'Steph
en Collinson', 'Pam Key, Pam Key', 'Peter Aitken', 'Elaine Mallon, Elaine Mallon',
'Neil Munro, Neil Munro', 'Robert Barnes', 'Chad Pergram', 'Chris Pandolfo', 'Deirdr
e Reilly', 'Katherine Hamilton, Katherine Hamilton', 'Brian Fung', 'Matt Egan', 'Joe
l B. Pollak, Joel B. Pollak', 'Elizabeth Elkind', 'Greg Norman', 'Peter Caddle, Pete
r Caddle', 'Frances Martel, Frances Martel', 'Haley Chi-Sing', 'Madeline Coggins',
'Kurt Zindulka, Kurt Zindulka', 'Rebecca Rosenberg', 'Houston Keene', 'Paulina Deda
j', 'Christian K. Caruzo, Christian K. Caruzo', 'Hannah Ray Lambert', 'AWR Hawkins,
AWR Hawkins', 'Andrew Miller', 'Nadeen Ebrahim', 'Danielle Wallace', 'Eric Bradner',
'Alana Mastrangelo, Alana Mastrangelo', 'Paul Steinhauser', 'Jordan Dixon-Hamilton,
Jordan Dixon-Hamilton', 'Paul Kane', 'Sean Moran, Sean Moran', 'Brianna Herlihy', 'G
lenn Kessler', 'Adam Shaw', 'Wendell Husebø, Wendell Husebø', 'Elizabeth Heckman',
'Hannah Rabinowitz', 'Joshua Nelson', 'Jessica Chasmar', 'John Nolte, John Nolte',
'Lindsay Kornick', 'Ryan Gaydos', 'Tierney Sneed', 'John Hayward, John Hayward', 'Az
i Paybarah', 'Zachary B. Wolf', 'Hannah Bleau, Hannah Bleau', 'Ashley Oliver, Ashley
Oliver', 'Joe Schoffstall', 'Nicole Goodkind', 'Emma Colton', 'Lucas Nolan, Lucas No
lan', 'Jeffrey Clark', 'Kevin Liptak', 'Melissa Rudy', 'Oliver Darcy', 'Chris Eberha
rt', 'Alexandra Meeks', 'Audrey Conklin', 'Fox News', 'Allum Bokhari, Allum Bokhar
i', 'Philip Bump', 'Steve Contorno', 'Kassy Dillon', 'Taylor Penley', 'Spencer S. Hs
u', 'Jon Brown', 'Kendall Tietz', 'Bailee Hill', 'Kurt Knutsson, CyberGuy Report',
'Adam Sabes', 'Paul Waldman', 'Ariane de Vogue', 'Michael Lee', 'Bob Price, Bob Pric
e', 'Reuters', 'Alisha Ebrahimji', 'Maeve Reston', 'Ashley Carnahan', 'Charles Creit
z', 'Peter Kasperowicz', 'Andrea Vacchiano', 'Angelica Stabile', 'Michael Ruiz', 'ms
n', 'Tony Romm', 'Louis Casiano', 'Anders Hagstrom', 'Kristine Parks', 'Joshua Klei
n, Joshua Klein', 'Warner Todd Huston, Warner Todd Huston', 'Sarah Rumpf-Whitten',
'David Ng, David Ng', 'Jacob Bliss, Jacob Bliss', 'Kyle Morris', 'Oliver JJ Lane, Ol
iver JJ Lane', 'Caitlin McFall', 'Yael Halon', 'John Binder, John Binder', 'Jeff Ste
in', 'Tami Luhby', 'Sean Lyngaas', 'Chantz Martin', 'Aaron Blake', 'John Wagner', 'P
atrick Hauf', 'Fox News Staff', 'Kerry Byrne', 'Brian Flood', 'Brie Stimson', 'Matth
ew Boyle, Matthew Boyle', 'Brandon Gillespie', 'Amy Furr, Amy Furr', 'Stephen Sorac
e', 'Aubrie Spady', 'Jeff Poor, Jeff Poor', 'Simon Kent, Simon Kent', 'Jennifer Rubi
n', 'Brooke Singman', 'Kristina Wong, Kristina Wong', 'Thomas Catenacci', 'Mariana A
lfaro', 'Joseph Wulfsohn', 'Ian Hanchett, Ian Hanchett', 'Aaron Kliegman', 'Greg Weh
ner', 'Nick Gilbertson, Nick Gilbertson', 'Hannah Grossman', 'Landon Mion', 'Dylan G
winn, Dylan Gwinn', 'Alexander Hall', 'Hannah Knowles', 'Paul Bois, Paul Bois'}
```

Assign class based on `source_name` and AllSides Media Bias Chart

In [14]:

```
slct_tbl_full_df03 = slct_tbl_full_df02[~slct_tbl_full_df02['author']\
                                        .isin(['msn',
                                               'Associated Press',
                                               'Reuters'])]
slct_tbl_full_df03 = slct_tbl_full_df03.reset_index()
slct_tbl_full_df03['political_lean'] = 'right'
print(slct_tbl_full_df03.shape)
display(slct_tbl_full_df03.head())
slct_tbl_full_df03.loc[(slct_tbl_full_df03['source_name'] \
                        == 'The Washington Post') \
                       | (slct_tbl_full_df03['source_name'] \
                          == 'CNN'), 'political_lean'] = 'left'
```

```
display(slct_tbl_full_df03.head())

display(slct_tbl_full_df03['political_lean'].value_counts())
```

```
(4026, 11)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

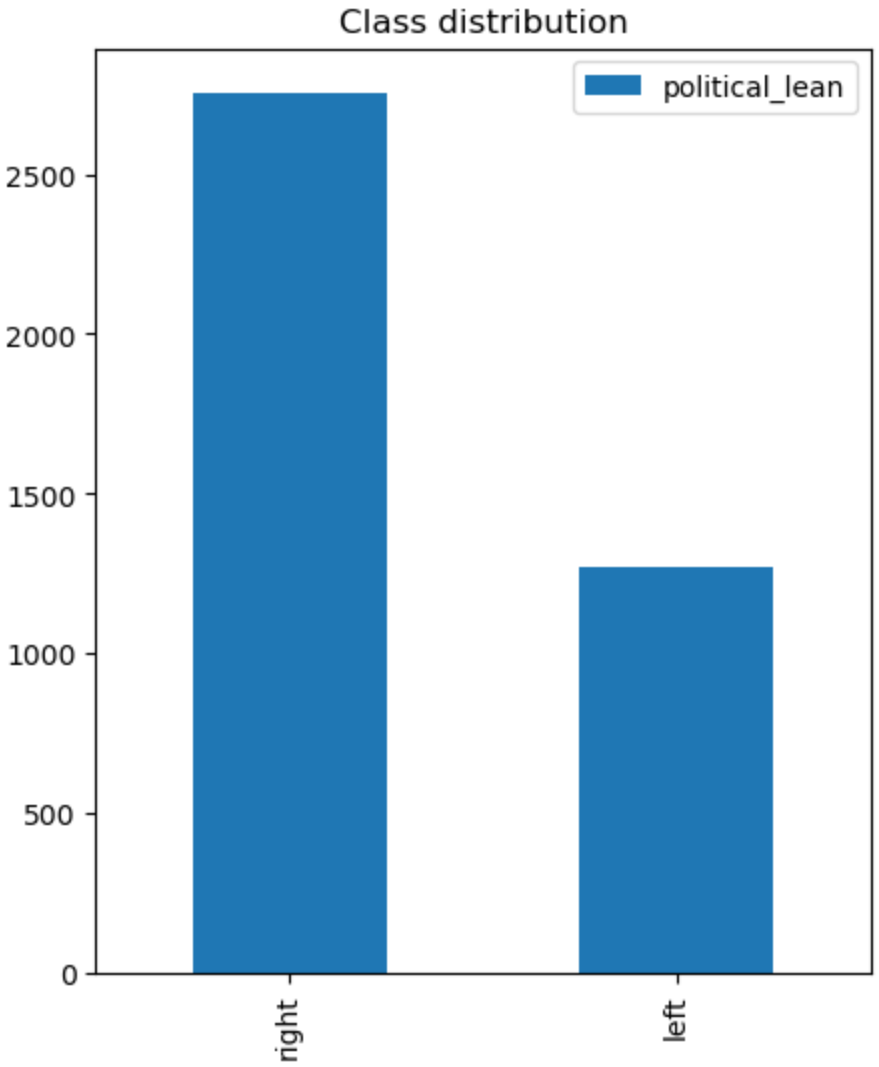
```
right    2758
left     1268
```

Visualize class distribution

There is definitely an imbalance in the number of instances in each class. This is due to Fox News being the most prolific source, whether because they put out a lot more articles or their sites were more consistently available for scraping. This imbalance is not considered extreme and will not be adjusted for within the scope of the current study.

In [15]:
```
slct_tbl_full_df03['political_lean'].value_counts().plot(kind="bar",
                                                         legend=True,
                                                         figsize=(5,6),
                                                        title='Class distribution')
```

Out[15]:
```
<Axes: title={'center': 'Class distribution'}>
```

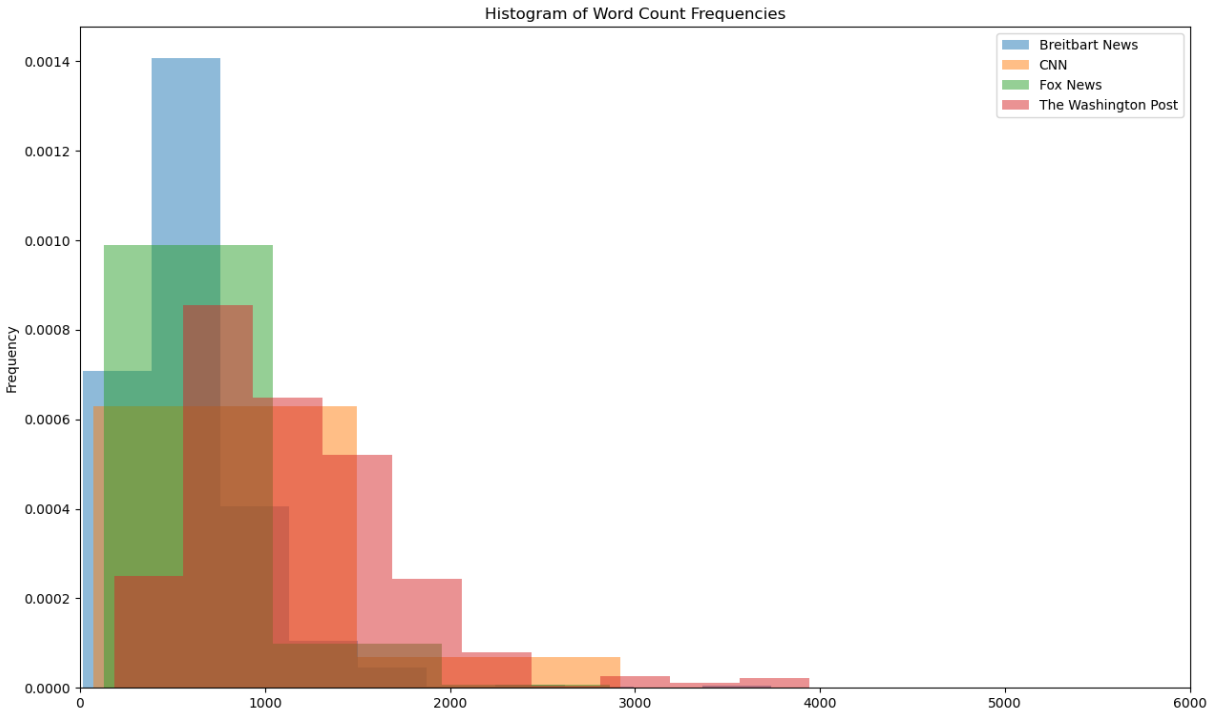


Plot word counts

All sources seem to have very similar consolidation of most frequent word counts between 0 and 2,000. However, the two "left" sources (CNN and The Washington Post) seem to be the significant source of the outliers, with a small amount of articles each that have extremely large word counts (*note:* the x-axis range was truncated at 6,000 to make it more readabla--as noted above, there were some articles with word counts greater than 14,000). Given the simialirties between the sources within each class, the differences may correlate to intentional word limitation based on perceived audience desires, but in the very least do add evidence that the sources have been grouped together appropriately.

In [16]:
```
slct_tbl_full_df03.groupby('source_name')['word_cnt'].plot(kind="hist",
                                                           density=True,
                                                           alpha=0.5,
                                                           legend=True,
                                                           figsize=(15,9),
                                    title='Histogram of Word Count Frequencies',
                                                           xlim=(0,6000))
```

Out[16]:
```
source_name
Breitbart News         Axes(0.125,0.11;0.775x0.77)
CNN                    Axes(0.125,0.11;0.775x0.77)
Fox News               Axes(0.125,0.11;0.775x0.77)
The Washington Post    Axes(0.125,0.11;0.775x0.77)
Name: word_cnt, dtype: object
```



Data preprocessing

In [17]:
```
def uniq_tok(df_col=None):
    '''Display all unique tokens across all instances'''
    df_cols1 = pd.Series(df_col)

    all_tokens_lst01 = []

    [all_tokens_lst01.append(f) for f in df_cols1]
    all_tokens_lst01 = list(itertools.chain.from_iterable(all_tokens_lst01))
    all_tokens_set01 = set(all_tokens_lst01)
    print(len(sorted(all_tokens_set01)))
    print(sorted(all_tokens_set01))
```

In [18]:
```
slct_tbl_full_df04 = slct_tbl_full_df03.copy()
```

Case-loading

In [19]:
```
slct_tbl_full_df03['lower'] = slct_tbl_full_df03['article_text']\
.apply(str.lower)

print(slct_tbl_full_df03.shape)
display(slct_tbl_full_df03.head())
```

```
(4026, 12)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

Text normalization

Create function

In [20]:
```
def normalize(text):
    text = tprep.normalize.hyphenated_words(text)
    text = tprep.normalize.quotation_marks(text)
    text = tprep.normalize.unicode(text)
    text = tprep.remove.accents(text)
    return text
```

Call function

In [21]:
```
slct_tbl_full_df03['norm'] = slct_tbl_full_df03['lower'].apply(normalize)

print(slct_tbl_full_df03.shape)
display(slct_tbl_full_df03.head())
```

```
for c in range(0,1):
    try:
        print(slct_tbl_full_df03['norm'][c], '\n')
    except:
        print(f'Skip {c}')
```

```
(4026, 13)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

travelers in alabama driving on interstate 65 to parties and barbecues on memorial d
ay might have seen messages on digital road signs honoring veterans who died fightin
g for the united states. but that's not what some drivers near clanton, ala., saw on
monday. instead, motorists reported seeing a sign that was apparently hacked to disp
lay the words "reclaim america," a white nationalist slogan, and "patriot front us,"
referencing the white supremacist group that was involved in the deadly 2017 unite t
he right rally in charlottesville. "how does this come about?" wrote sarah hughes, a
motorist who captured photos of the sign and posted them on twitter. "weird as hel
l." a contractor's portable message board was hacked on i-65 in chilton county, al
a., on monday afternoon, john mcwilliams, a spokesman for the alabama department of
transportation (aldot) west central region, told the washington post in a statement.
"a citizen alerted a nearby state trooper about the message, who then contacted aldo
t," mcwilliams said tuesday. "aldot personnel immediately responded and turned the m
essage board off. no other message boards on i-65 were affected." mcwilliams added
that aldot is investigating how the white supremacist language appeared on the sign
near clanton, about 40 miles northwest of montgomery, ala. officials have given no i
mmediate indication of who is responsible for apparently hacking the interstate sig
n. the news was first reported by al.com. hughes told the post that she was driving
home to birmingham from a weekend at alabama's gulf coast when she saw the white sup
remacist messages that have recently popped up around her home city from supporters
of patriot front. "when i saw it, i thought, 'oh, it's the same guys,' " said hughe
s, a 31-year-old attorney. "i was kind of shocked." the hacked alabama road sign com
es at a time when president biden has declared white supremacy "the most dangerous t
errorist threat" to the country. during his commencement address at howard universit
y this month, biden told the graduating class at the historically black university t
hat he pledged "to stand up against the poison of white supremacy, as i did in my in
augural address — to single it out as the most dangerous terrorist threat to our hom
eland is white supremacy." "i don't have to tell you that progress toward justice of
ten meets ferocious pushback from the oldest and most sinister of forces," biden sai
d in the may 13 address, after quoting donald trump's equivocating response to the 2
017 rally in charlottesville that killed 32-year-old heather heyer and injured 19 ot
hers. "that's because hate never goes away." biden calls white supremacy greatest te
rrorism threat as 2024 race heats up the southern poverty law center (splc) tracked
at least 13 hate groups in alabama in 2021, including the proud boys. the discussion
surrounding white supremacists and white nationalists in alabama intensified this mo
nth after sen. tommy tuberville (r-ala.) said that people identified as "white extre
mists" and white nationalists should be allowed to serve in the u.s. armed forces. w
hen asked by a reporter with wbhm in birmingham whether white nationalists should be
allowed to serve in the military, tuberville replied, "well, they call them that. i
call them americans." after tuberville was criticized, a spokesman told the post tha
t the senator "resents the implication that the people in our military are anything
but patriots and heroes." gop senator says of white nationalists in the military, 'i
call them americans' patriot front, the white supremacist group whose name was displ
ayed on the interstate sign, is a texas-based hate group that broke off from vanguar
d america and formed after the charlottesville rally, the splc says. its members hav
e chanted "reclaim america" at rallies in coeur d'alene, idaho, washington and bosto
n in recent years, according to news reports. patriot front is responsible for "the
vast majority of white supremacist propaganda distributed in the united states" sinc
e 2019, according to the anti-defamation league. it's not the first time that langua
ge promoting patriot front has made its way into a public space in alabama. in july,
graffiti beneath a birmingham bridge appeared with "patriot front us" spray-painted
in red and blue letters, al.com reported. other patriot front graffiti has also been
spotted in birmingham, a city with a population that's nearly 70 percent black, acco
rding to u.s. census data. a photo posted to twitter this month showed more patriot
front graffiti along the red mountain expressway in birmingham with the words, "we d
are defend our rights." the patriot front graffiti was later removed, but the messag

e left sydney duncan, the attorney director for the magic legal center in birmingha
m, saddened that hate had become so public in some parts of alabama. "white supremac
y is alive and well," duncan wrote. hughes said she was traveling north to birmingh
am when she pulled over on i-65 to take photos of the messages on the sign. she had
seen confederate monuments and flags on that drive before, but that kind of messagin
g on government-owned property was different, she said. a police officer who was alr
eady at the scene waved at her to keep driving, hughes added. when she returned hom
e, hughes said she felt compelled to share the images due to the ongoing conversatio
n happening among birmingham residents about the promotion of patriot front in publi
c spaces. "some people might perceive this as upsetting and scary, and a sign of the
worsening of our country," she said. "but if this is their strategy, then i'm not re
ally impressed." she added, "they're a dying breed." toluse olorunnipa and azi payba
rah contributed to this report.

In [22]:
```
text2find_rex = rex.compile(r'(click here to get the fox news app)')
test_lst = []

def test(text):
    test_lst.append(text2find_rex.findall(text))

slct_tbl_full_df03['norm'].apply(test)

display(slct_tbl_full_df03.head())
#print(test_lst)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

Remove special characters

Create function

In [23]:
```
rex_sep = rex.compile(r' ')
rex_ucode = rex.compile(r'[\\]u20*')

'''re.sub lambda citation:
https://chat.openai.com/share/402ec66e-2802-4cda-af8c-6f9f5b097d85
'''
sep_lst = []
ucode_lst = []
# Add leading and trailing space to URLs
def rex_replace(text):
    #txt = str(text)
    #print(lambda x: x.replace(' ', ' '))
    #sep_lst.append(rex_sep.findall(txt))
    #ucode_lst.append(rex_ucode.findall(txt))
```

```
    text = text.replace(r' ', ' ').replace(r'-', ' ')\
    .replace(r'\n', ' ').replace('\u2063', ' ').replace('\u2066', ' ')\
    .replace('\u2069', ' ').replace('\u200b', ' ').replace('\u200d', ' ')\
    .replace('(click to view)', ' ')\
    .replace('a post shared by', ' ')\
    .replace('app users click here', ' ')\
    .replace('app users: click here', ' ')\
    .replace('app users, click here:', ' ')\
    .replace('click here.', ' ')\
    .replace('click here for more cartoons', ' ')\
    .replace('click here for more', ' ')\
    .replace('click here for more sports coverage on foxnews.com', ' ')\
    .replace('click here for other fox news digital adoptable pets stories', ' ')\
    .replace('click here for the fox news app', ' ')\
    .replace('click here for the latest fox news reporting', ' ')\
    .replace('click here for topline and cross tabs conducted', ' ')\
    .replace('click here to hear more', ' ')\
    .replace('click here to ge the fox news app', ' ')\
    .replace('click here to get the fox news app', ' ')\
    .replace('click here to get the opinion newsletter', ' ')\
    .replace('click here to learn more', ' ')\
    .replace('click here to read more', ' ')\
    .replace('click here to sign up for our health newsletter', ' ')\
    .replace('click here to sign up for our lifestyle newsletter', ' ')\
    .replace('click here to sign up for our opinion newsletter', ' ')\
    .replace('click here to sign up for the entertainment newsletter', ' ')\
    .replace('click here to subscribe and get your first year of fox nation free of
    .replace('click here to view', ' ')\
    .replace("click to get kurt's cyberguy newsletter with quick tips, tech reviews
    .replace("click to get kurt's cyberguy newsletter with security alerts, quick t
    .replace("click to get kurt's free cyberguy newsletter with quick tips, tech re
    .replace("click to get kurt's free cyberguy newsletter with security alerts, qu
    .replace('click to get the fox news app', ' ')\
    .replace('fox news digital', ' ')\
    .replace('request for comment', ' ')\
    .replace('the ap ', ' ')\
    .replace('copyright © 2023 breitbart', ' ')\
    .replace('all rights reserved', ' ')\
    .replace('copyright 2023 cyberguy.com', ' ')\
    .replace('copyright 2023 fox news network', ' ')\
    .replace('copyright 2023 viq media transcription', ' ')\
    .replace("please let us know if you're having issues with commenting", ' ')\
    .replace('view this post on instagram', ' ')
    #txt = txt
    #text = text.replace(r'200b', 'd171c')
    #text = rex_ucode.sub('', text)
    return text

    #.replace('philip bump', ' ')\
    #.replace('paul kane', ' ')\
    #.replace('&amp', ' ')\
```

Call function

In [24]:
```
slct_tbl_full_df03['replace'] = slct_tbl_full_df03['norm'].apply(rex_replace)

#print(ucode_lst)
#print(sep_lst)

print(slct_tbl_full_df03.shape)
display(slct_tbl_full_df03.head())
```

```
(4026, 14)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

'''Complex citation (add lambda): https://chat.openai.com/share/a135754c-c38c-47ea-8f83-54d41d5397ab ''' slct_tbl_full_df03['replace'] = slct_tbl_full_df03['norm'].apply(lambda x: x.replace(' ', ' ').replace(r'\n', ' ').replace('\u2063', ' ').replace('\u2066', ' ').replace('\u2069', ' ').replace('\u200b', ' ').replace('\u200d', ' '))

URL RegEx find

Create function

In [25]:
```
rex_url_c = rex.compile(r'http[s]?:[\/]+[\S]*\s')

'''re.sub lambda citation:
https://chat.openai.com/share/402ec66e-2802-4cda-af8c-6f9f5b097d85
'''
# Add leading and trailing space to URLs
def rex_url(text):
    text = rex_url_c.sub(lambda match: ' ' + match.group(0) + ' ', text)
    return text
```

Call function

In [26]:
```
slct_tbl_full_df03['rex_urls'] = slct_tbl_full_df03['replace'].apply(rex_url)

print(slct_tbl_full_df03.shape)
display(slct_tbl_full_df03.head())
```

```
(4026, 15)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

Separate emojis as individual tokens

Create function

In [27]:
```
def emoji_split(text):
    return("".join([' ' + c + ' ' if emoji.is_emoji(c) else c for c in text]))
```

Call function

In [28]:
```
slct_tbl_full_df03['emoji_split'] = slct_tbl_full_df03['rex_urls']\
.apply(emoji_split)

print(slct_tbl_full_df03.shape)
display(slct_tbl_full_df03.head())

for c in range(0,1):
    try:
        print(slct_tbl_full_df03['emoji_split'][c], '\n')
    except:
        print(f'Skip {c}')
```

```
(4026, 16)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

travelers in alabama driving on interstate 65 to parties and barbecues on memorial d
ay might have seen messages on digital road signs honoring veterans who died fightin
g for the united states. but that's not what some drivers near clanton, ala., saw on
monday. instead, motorists reported seeing a sign that was apparently hacked to disp
lay the words "reclaim america," a white nationalist slogan, and "patriot front us,"
referencing the white supremacist group that was involved in the deadly 2017 unite t
he right rally in charlottesville. "how does this come about?" wrote sarah hughes, a
motorist who captured photos of the sign and posted them on twitter. "weird as hel
l." a contractor's portable message board was hacked on i 65 in chilton county, al
a., on monday afternoon, john mcwilliams, a spokesman for the alabama department of
transportation (aldot) west central region, told the washington post in a statement.
"a citizen alerted a nearby state trooper about the message, who then contacted aldo
t," mcwilliams said tuesday. "aldot personnel immediately responded and turned the m
essage board off. no other message boards on i 65 were affected." mcwilliams added
that aldot is investigating how the white supremacist language appeared on the sign
near clanton, about 40 miles northwest of montgomery, ala. officials have given no i
mmediate indication of who is responsible for apparently hacking the interstate sig
n. the news was first reported by al.com. hughes told the post that she was driving
home to birmingham from a weekend at alabama's gulf coast when she saw the white sup
remacist messages that have recently popped up around her home city from supporters
of patriot front. "when i saw it, i thought, 'oh, it's the same guys,' " said hughe
s, a 31 year old attorney. "i was kind of shocked." the hacked alabama road sign com
es at a time when president biden has declared white supremacy "the most dangerous t
errorist threat" to the country. during his commencement address at howard universit
y this month, biden told the graduating class at the historically black university t
hat he pledged "to stand up against the poison of white supremacy, as i did in my in
augural address — to single it out as the most dangerous terrorist threat to our hom
eland is white supremacy." "i don't have to tell you that progress toward justice of
ten meets ferocious pushback from the oldest and most sinister of forces," biden sai
d in the may 13 address, after quoting donald trump's equivocating response to the 2
017 rally in charlottesville that killed 32 year old heather heyer and injured 19 ot
hers. "that's because hate never goes away." biden calls white supremacy greatest te
rrorism threat as 2024 race heats up the southern poverty law center (splc) tracked
at least 13 hate groups in alabama in 2021, including the proud boys. the discussion
surrounding white supremacists and white nationalists in alabama intensified this mo
nth after sen. tommy tuberville (r ala.) said that people identified as "white extre
mists" and white nationalists should be allowed to serve in the u.s. armed forces. w
hen asked by a reporter with wbhm in birmingham whether white nationalists should be
allowed to serve in the military, tuberville replied, "well, they call them that. i
call them americans." after tuberville was criticized, a spokesman told the post tha
t the senator "resents the implication that the people in our military are anything
but patriots and heroes." gop senator says of white nationalists in the military, 'i
call them americans' patriot front, the white supremacist group whose name was displ
ayed on the interstate sign, is a texas based hate group that broke off from vanguar
d america and formed after the charlottesville rally, the splc says. its members hav
e chanted "reclaim america" at rallies in coeur d'alene, idaho, washington and bosto
n in recent years, according to news reports. patriot front is responsible for "the
vast majority of white supremacist propaganda distributed in the united states" sinc
e 2019, according to the anti defamation league. it's not the first time that langua
ge promoting patriot front has made its way into a public space in alabama. in july,
graffiti beneath a birmingham bridge appeared with "patriot front us" spray painted
in red and blue letters, al.com reported. other patriot front graffiti has also been
spotted in birmingham, a city with a population that's nearly 70 percent black, acco
rding to u.s. census data. a photo posted to twitter this month showed more patriot
front graffiti along the red mountain expressway in birmingham with the words, "we d
are defend our rights." the patriot front graffiti was later removed, but the messag

```
e left sydney duncan, the attorney director for the magic legal center in birmingha
m, saddened that hate had become so public in some parts of alabama. "white supremac
y is alive and well," duncan wrote.  hughes said she was traveling north to birmingh
am when she pulled over on i 65 to take photos of the messages on the sign. she had
seen confederate monuments and flags on that drive before, but that kind of messagin
g on government owned property was different, she said. a police officer who was alr
eady at the scene waved at her to keep driving, hughes added. when she returned hom
e, hughes said she felt compelled to share the images due to the ongoing conversatio
n happening among birmingham residents about the promotion of patriot front in publi
c spaces. "some people might perceive this as upsetting and scary, and a sign of the
worsening of our country," she said. "but if this is their strategy, then i'm not re
ally impressed." she added, "they're a dying breed." toluse olorunnipa and azi payba
rah contributed to this report.
```

Lemmatization using spaCY

In [29]:
```
nlp_trans = spacy.load('en_core_web_sm')

def lemma(text):
    trans_txt = nlp_trans(text)
    tokens = [t.lemma_ for t in trans_txt]
    return tokens
```

slct_tbl_full_df03['lemma'] = slct_tbl_full_df03['replace'].progress_apply(lemma) print(slct_tbl_full_df03.shape) display(slct_tbl_full_df03.head()) for c in range(0,1): try: print(slct_tbl_full_df03['lemma'][c], '\n') except: print(f'Skip {c}')

Display globally unique tokens on lemmas

In [30]:
```
#uniq_tok(df_col=slct_tbl_full_df03['lemma'])
```

Split text

Apply

In [31]:
```
slct_tbl_full_df03['split'] = slct_tbl_full_df03['emoji_split']\
.apply(str.split)

print(slct_tbl_full_df03.shape)
display(slct_tbl_full_df03.head())

for c in range(0,1):
    try:
        print(slct_tbl_full_df03['split'][c], '\n')
    except:
        print(f'Skip {c}')
```

```
(4026, 17)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

['travelers', 'in', 'alabama', 'driving', 'on', 'interstate', '65', 'to', 'parties',
'and', 'barbecues', 'on', 'memorial', 'day', 'might', 'have', 'seen', 'messages', 'o
n', 'digital', 'road', 'signs', 'honoring', 'veterans', 'who', 'died', 'fighting',
'for', 'the', 'united', 'states.', 'but', "that's", 'not', 'what', 'some', 'driver
s', 'near', 'clanton,', 'ala.,', 'saw', 'on', 'monday.', 'instead,', 'motorists', 'r
eported', 'seeing', 'a', 'sign', 'that', 'was', 'apparently', 'hacked', 'to', 'displ
ay', 'the', 'words', '"reclaim', 'america,"', 'a', 'white', 'nationalist', 'sloga
n,', 'and', '"patriot', 'front', 'us,"', 'referencing', 'the', 'white', 'supremacis
t', 'group', 'that', 'was', 'involved', 'in', 'the', 'deadly', '2017', 'unite', 'th
e', 'right', 'rally', 'in', 'charlottesville.', '"how', 'does', 'this', 'come', 'abo
ut?"', 'wrote', 'sarah', 'hughes,', 'a', 'motorist', 'who', 'captured', 'photos', 'o
f', 'the', 'sign', 'and', 'posted', 'them', 'on', 'twitter.', '"weird', 'as', 'hel
l."', 'a', "contractor's", 'portable', 'message', 'board', 'was', 'hacked', 'on',
'i', '65', 'in', 'chilton', 'county,', 'ala.,', 'on', 'monday', 'afternoon,', 'joh
n', 'mcwilliams,', 'a', 'spokesman', 'for', 'the', 'alabama', 'department', 'of', 't
ransportation', '(aldot)', 'west', 'central', 'region,', 'told', 'the', 'washingto
n', 'post', 'in', 'a', 'statement.', '"a', 'citizen', 'alerted', 'a', 'nearby', 'sta
te', 'trooper', 'about', 'the', 'message,', 'who', 'then', 'contacted', 'aldot,"',
'mcwilliams', 'said', 'tuesday.', '"aldot', 'personnel', 'immediately', 'responded',
'and', 'turned', 'the', 'message', 'board', 'off.', 'no', 'other', 'message', 'board
s', 'on', 'i', '65', 'were', 'affected."', 'mcwilliams', 'added', 'that', 'aldot',
'is', 'investigating', 'how', 'the', 'white', 'supremacist', 'language', 'appeared',
'on', 'the', 'sign', 'near', 'clanton,', 'about', '40', 'miles', 'northwest', 'of',
'montgomery,', 'ala.', 'officials', 'have', 'given', 'no', 'immediate', 'indicatio
n', 'of', 'who', 'is', 'responsible', 'for', 'apparently', 'hacking', 'the', 'inters
tate', 'sign.', 'the', 'news', 'was', 'first', 'reported', 'by', 'al.com.', 'hughe
s', 'told', 'the', 'post', 'that', 'she', 'was', 'driving', 'home', 'to', 'birmingha
m', 'from', 'a', 'weekend', 'at', "alabama's", 'gulf', 'coast', 'when', 'she', 'sa
w', 'the', 'white', 'supremacist', 'messages', 'that', 'have', 'recently', 'popped',
'up', 'around', 'her', 'home', 'city', 'from', 'supporters', 'of', 'patriot', 'fron
t.', '"when', 'i', 'saw', 'it,', 'i', 'thought,', "'oh,", "it's", 'the', 'same', "gu
ys,'", '"', 'said', 'hughes,', 'a', '31', 'year', 'old', 'attorney.', '"i', 'was',
'kind', 'of', 'shocked."', 'the', 'hacked', 'alabama', 'road', 'sign', 'comes', 'a
t', 'a', 'time', 'when', 'president', 'biden', 'has', 'declared', 'white', 'supremac
y', '"the', 'most', 'dangerous', 'terrorist', 'threat"', 'to', 'the', 'country.', 'd
uring', 'his', 'commencement', 'address', 'at', 'howard', 'university', 'this', 'mon
th,', 'biden', 'told', 'the', 'graduating', 'class', 'at', 'the', 'historically', 'b
lack', 'university', 'that', 'he', 'pledged', '"to', 'stand', 'up', 'against', 'th
e', 'poison', 'of', 'white', 'supremacy,', 'as', 'i', 'did', 'in', 'my', 'inaugura
l', 'address', '—', 'to', 'single', 'it', 'out', 'as', 'the', 'most', 'dangerous',
'terrorist', 'threat', 'to', 'our', 'homeland', 'is', 'white', 'supremacy."', '"i',
"don't", 'have', 'to', 'tell', 'you', 'that', 'progress', 'toward', 'justice', 'ofte
n', 'meets', 'ferocious', 'pushback', 'from', 'the', 'oldest', 'and', 'most', 'sinis
ter', 'of', 'forces,"', 'biden', 'said', 'in', 'the', 'may', '13', 'address,', 'afte
r', 'quoting', 'donald', "trump's", 'equivocating', 'response', 'to', 'the', '2017',
'rally', 'in', 'charlottesville', 'that', 'killed', '32', 'year', 'old', 'heather',
'heyer', 'and', 'injured', '19', 'others.', '"that\'s', 'because', 'hate', 'never',
'goes', 'away."', 'biden', 'calls', 'white', 'supremacy', 'greatest', 'terrorism',
'threat', 'as', '2024', 'race', 'heats', 'up', 'the', 'southern', 'poverty', 'law',
'center', '(splc)', 'tracked', 'at', 'least', '13', 'hate', 'groups', 'in', 'alabam
a', 'in', '2021,', 'including', 'the', 'proud', 'boys.', 'the', 'discussion', 'surro
unding', 'white', 'supremacists', 'and', 'white', 'nationalists', 'in', 'alabama',
'intensified', 'this', 'month', 'after', 'sen.', 'tommy', 'tuberville', '(r', 'al
a.)', 'said', 'that', 'people', 'identified', 'as', '"white', 'extremists"', 'and',
'white', 'nationalists', 'should', 'be', 'allowed', 'to', 'serve', 'in', 'the', 'u.
s.', 'armed', 'forces.', 'when', 'asked', 'by', 'a', 'reporter', 'with', 'wbhm', 'i

```
n', 'birmingham', 'whether', 'white', 'nationalists', 'should', 'be', 'allowed', 't
o', 'serve', 'in', 'the', 'military,', 'tuberville', 'replied,', '"well,', 'they',
'call', 'them', 'that.', 'i', 'call', 'them', 'americans."', 'after', 'tuberville',
'was', 'criticized,', 'a', 'spokesman', 'told', 'the', 'post', 'that', 'the', 'senat
or', '"resents', 'the', 'implication', 'that', 'the', 'people', 'in', 'our', 'milita
ry', 'are', 'anything', 'but', 'patriots', 'and', 'heroes."', 'gop', 'senator', 'say
s', 'of', 'white', 'nationalists', 'in', 'the', 'military,', "'i", 'call', 'them',
"americans'", 'patriot', 'front,', 'the', 'white', 'supremacist', 'group', 'whose',
'name', 'was', 'displayed', 'on', 'the', 'interstate', 'sign,', 'is', 'a', 'texas',
'based', 'hate', 'group', 'that', 'broke', 'off', 'from', 'vanguard', 'america', 'an
d', 'formed', 'after', 'the', 'charlottesville', 'rally,', 'the', 'splc', 'says.',
'its', 'members', 'have', 'chanted', '"reclaim', 'america"', 'at', 'rallies', 'in',
'coeur', "d'alene,", 'idaho,', 'washington', 'and', 'boston', 'in', 'recent', 'year
s,', 'according', 'to', 'news', 'reports.', 'patriot', 'front', 'is', 'responsible',
'for', '"the', 'vast', 'majority', 'of', 'white', 'supremacist', 'propaganda', 'dist
ributed', 'in', 'the', 'united', 'states"', 'since', '2019,', 'according', 'to', 'th
e', 'anti', 'defamation', 'league.', "it's", 'not', 'the', 'first', 'time', 'that',
'language', 'promoting', 'patriot', 'front', 'has', 'made', 'its', 'way', 'into',
'a', 'public', 'space', 'in', 'alabama.', 'in', 'july,', 'graffiti', 'beneath', 'a',
'birmingham', 'bridge', 'appeared', 'with', '"patriot', 'front', 'us"', 'spray', 'pa
inted', 'in', 'red', 'and', 'blue', 'letters,', 'al.com', 'reported.', 'other', 'pat
riot', 'front', 'graffiti', 'has', 'also', 'been', 'spotted', 'in', 'birmingham,',
'a', 'city', 'with', 'a', 'population', "that's", 'nearly', '70', 'percent', 'blac
k,', 'according', 'to', 'u.s.', 'census', 'data.', 'a', 'photo', 'posted', 'to', 'tw
itter', 'this', 'month', 'showed', 'more', 'patriot', 'front', 'graffiti', 'along',
'the', 'red', 'mountain', 'expressway', 'in', 'birmingham', 'with', 'the', 'words,',
'"we', 'dare', 'defend', 'our', 'rights."', 'the', 'patriot', 'front', 'graffiti',
'was', 'later', 'removed,', 'but', 'the', 'message', 'left', 'sydney', 'duncan,', 't
he', 'attorney', 'director', 'for', 'the', 'magic', 'legal', 'center', 'in', 'birmin
gham,', 'saddened', 'that', 'hate', 'had', 'become', 'so', 'public', 'in', 'some',
'parts', 'of', 'alabama.', '"white', 'supremacy', 'is', 'alive', 'and', 'well,"', 'd
uncan', 'wrote.', 'hughes', 'said', 'she', 'was', 'traveling', 'north', 'to', 'birmi
ngham', 'when', 'she', 'pulled', 'over', 'on', 'i', '65', 'to', 'take', 'photos', 'o
f', 'the', 'messages', 'on', 'the', 'sign.', 'she', 'had', 'seen', 'confederate', 'm
onuments', 'and', 'flags', 'on', 'that', 'drive', 'before,', 'but', 'that', 'kind',
'of', 'messaging', 'on', 'government', 'owned', 'property', 'was', 'different,', 'sh
e', 'said.', 'a', 'police', 'officer', 'who', 'was', 'already', 'at', 'the', 'scen
e', 'waved', 'at', 'her', 'to', 'keep', 'driving,', 'hughes', 'added.', 'when', 'sh
e', 'returned', 'home,', 'hughes', 'said', 'she', 'felt', 'compelled', 'to', 'shar
e', 'the', 'images', 'due', 'to', 'the', 'ongoing', 'conversation', 'happening', 'am
ong', 'birmingham', 'residents', 'about', 'the', 'promotion', 'of', 'patriot', 'fron
t', 'in', 'public', 'spaces.', '"some', 'people', 'might', 'perceive', 'this', 'as',
'upsetting', 'and', 'scary,', 'and', 'a', 'sign', 'of', 'the', 'worsening', 'of', 'o
ur', 'country,"', 'she', 'said.', '"but', 'if', 'this', 'is', 'their', 'strategy,',
'then', "i'm", 'not', 'really', 'impressed."', 'she', 'added,', '"they\'re', 'a', 'd
ying', 'breed."', 'toluse', 'olorunnipa', 'and', 'azi', 'paybarah', 'contributed',
'to', 'this', 'report.']
```

Display globally unique tokens on first split

In [32]:
```
#uniq_tok(df_col=slct_tbl_full_df03['split'])
```

Remove stop words

In [33]:
```
sw = stopwords.words("english")

# Add additional stop words
sw.extend(['',
           '',
           'arent',
           'cannot',
           'cant',
           'couldnt',
           'couldve',
           'didnt',
           'doesnt',
           'dont',
           'hadnt',
           'hasnt',
           'havent',
           'hes',
           'im',
           "i'm",
           'isnt',
           'it’s',
           'ive',
           'of',
           'mightnt',
           'mustnt',
           'neednt',
           'shant',
           'shes',
           'shouldnt',
           'shouldve',
           'thatll',
           'theyll',
           'theyve',
           'wasnt',
           'werent',
           'whats',
           'weve',
           'wont',
           'wouldnt',
           'wouldve',
           'yall',
           'youd',
           'youll',
           'youre',
           'youve',
           "we'll",
           "you’re",
           "you’ve",
           "you’ll",
           "you’d",
           "she’s",
           "it’s",
           "that’ll",
           "don’t",
           "should’ve",
```

```
            "aren't",
            "couldn't",
            "didn't",
            "doesn't",
            "hadn't",
            "hasn't",
            "haven't",
            "isn't",
            "mightn't",
            "mustn't",
            "needn't",
            "shan't",
            "shouldn't",
            "wasn't",
            "weren't",
            "won't",
            "wouldn't",
            "i'm",
            "we'll",
            'said',
            'told',
            'according',
            'fox',
            'news',
            'cnn',
            'breitbart',
            'reuters',
            'reporting',
            'reported',
            #'statement',
            #'spoke',
            #'next',
            #'though',
            #'often',
            #'story',
            #'updated',
            #'additional',
            #'developments',
            #'follow',
           ])

print(sw)
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "yo
u've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'h
is', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itse
lf', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'who
m', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were',
'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing',
'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of',
'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'durin
g', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'o
n', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'wh
en', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'ot
her', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'to
o', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've",
'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "cou
ldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'h
aven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'n
eedn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'were
n', "weren't", 'won', "won't", 'wouldn', "wouldn't", '', '', 'arent', 'cannot', 'can
t', 'couldnt', 'couldve', 'didnt', 'doesnt', 'dont', 'hadnt', 'hasnt', 'havent', 'he
s', 'im', "i'm", 'isnt', 'it’s', 'ive', 'of', 'mightnt', 'mustnt', 'neednt', 'shan
t', 'shes', 'shouldnt', 'shouldve', 'thatll', 'theyll', 'theyve', 'wasnt', 'werent',
'whats', 'weve', 'wont', 'wouldnt', 'wouldve', 'yall', 'youd', 'youll', 'youre', 'yo
uve', "we'll", 'you’re', 'you’ve', 'you’ll', 'you’d', 'she’s', 'it’s', 'that’ll', 'd
on’t', 'should’ve', 'aren’t', 'couldn’t', 'didn’t', 'doesn’t', 'hadn’t', 'hasn’t',
'haven’t', 'isn’t', 'mightn’t', 'mustn’t', 'needn’t', 'shan’t', 'shouldn’t', 'was
n’t', 'weren’t', 'won’t', 'wouldn’t', 'i’m', 'we’ll', 'said', 'told', 'according',
'fox', 'news', 'cnn', 'breitbart', 'reuters', 'reporting', 'reported']
```

Create function

In [34]:
```
def sw_remover(tokens):
    return [t for t in tokens if t.lower() not in sw]
```

Call function

In [35]:
```
slct_tbl_full_df03['no_sw'] = slct_tbl_full_df03['split'].apply(sw_remover)

print(slct_tbl_full_df03.shape)
display(slct_tbl_full_df03.head())

for c in range(0,1):
    print(slct_tbl_full_df03['no_sw'][c])
```

```
(4026, 18)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

['travelers', 'alabama', 'driving', 'interstate', '65', 'parties', 'barbecues', 'mem
orial', 'day', 'might', 'seen', 'messages', 'digital', 'road', 'signs', 'honoring',
'veterans', 'died', 'fighting', 'united', 'states.', "that's", 'drivers', 'near', 'c
lanton,', 'ala.,', 'saw', 'monday.', 'instead,', 'motorists', 'seeing', 'sign', 'app
arently', 'hacked', 'display', 'words', '"reclaim', 'america,"', 'white', 'nationali
st', 'slogan,', '"patriot', 'front', 'us,"', 'referencing', 'white', 'supremacist',
'group', 'involved', 'deadly', '2017', 'unite', 'right', 'rally', 'charlottesvill
e.', '"how', 'come', 'about?"', 'wrote', 'sarah', 'hughes,', 'motorist', 'captured',
'photos', 'sign', 'posted', 'twitter.', '"weird', 'hell."', "contractor's", 'portabl
e', 'message', 'board', 'hacked', '65', 'chilton', 'county,', 'ala.,', 'monday', 'af
ternoon,', 'john', 'mcwilliams,', 'spokesman', 'alabama', 'department', 'transportat
ion', '(aldot)', 'west', 'central', 'region,', 'washington', 'post', 'statement.',
'"a', 'citizen', 'alerted', 'nearby', 'state', 'trooper', 'message,', 'contacted',
'aldot,"', 'mcwilliams', 'tuesday.', '"aldot', 'personnel', 'immediately', 'responde
d', 'turned', 'message', 'board', 'off.', 'message', 'boards', '65', 'affected."',
'mcwilliams', 'added', 'aldot', 'investigating', 'white', 'supremacist', 'language',
'appeared', 'sign', 'near', 'clanton,', '40', 'miles', 'northwest', 'montgomery,',
'ala.', 'officials', 'given', 'immediate', 'indication', 'responsible', 'apparentl
y', 'hacking', 'interstate', 'sign.', 'first', 'al.com.', 'hughes', 'post', 'drivin
g', 'home', 'birmingham', 'weekend', "alabama's", 'gulf', 'coast', 'saw', 'white',
'supremacist', 'messages', 'recently', 'popped', 'around', 'home', 'city', 'supporte
rs', 'patriot', 'front.', '"when', 'saw', 'it,', 'thought,', "'oh,", "guys,'", '"',
'hughes,', '31', 'year', 'old', 'attorney.', '"i', 'kind', 'shocked."', 'hacked', 'a
labama', 'road', 'sign', 'comes', 'time', 'president', 'biden', 'declared', 'white',
'supremacy', '"the', 'dangerous', 'terrorist', 'threat"', 'country.', 'commencemen
t', 'address', 'howard', 'university', 'month,', 'biden', 'graduating', 'class', 'hi
storically', 'black', 'university', 'pledged', '"to', 'stand', 'poison', 'white', 's
upremacy,', 'inaugural', 'address', '—', 'single', 'dangerous', 'terrorist', 'threa
t', 'homeland', 'white', 'supremacy."', '"i', 'tell', 'progress', 'toward', 'justic
e', 'often', 'meets', 'ferocious', 'pushback', 'oldest', 'sinister', 'forces,"', 'bi
den', 'may', '13', 'address,', 'quoting', 'donald', "trump's", 'equivocating', 'resp
onse', '2017', 'rally', 'charlottesville', 'killed', '32', 'year', 'old', 'heather',
'heyer', 'injured', '19', 'others.', '"that\'s', 'hate', 'never', 'goes', 'away."',
'biden', 'calls', 'white', 'supremacy', 'greatest', 'terrorism', 'threat', '2024',
'race', 'heats', 'southern', 'poverty', 'law', 'center', '(splc)', 'tracked', 'leas
t', '13', 'hate', 'groups', 'alabama', '2021,', 'including', 'proud', 'boys.', 'disc
ussion', 'surrounding', 'white', 'supremacists', 'white', 'nationalists', 'alabama',
'intensified', 'month', 'sen.', 'tommy', 'tuberville', '(r', 'ala.)', 'people', 'ide
ntified', '"white', 'extremists"', 'white', 'nationalists', 'allowed', 'serve', 'u.
s.', 'armed', 'forces.', 'asked', 'reporter', 'wbhm', 'birmingham', 'whether', 'whit
e', 'nationalists', 'allowed', 'serve', 'military,', 'tuberville', 'replied,', '"wel
l,', 'call', 'that.', 'call', 'americans."', 'tuberville', 'criticized,', 'spokesma
n', 'post', 'senator', '"resents', 'implication', 'people', 'military', 'anything',
'patriots', 'heroes."', 'gop', 'senator', 'says', 'white', 'nationalists', 'militar
y,', "'i", 'call', "americans'", 'patriot', 'front,', 'white', 'supremacist', 'grou
p', 'whose', 'name', 'displayed', 'interstate', 'sign,', 'texas', 'based', 'hate',
'group', 'broke', 'vanguard', 'america', 'formed', 'charlottesville', 'rally,', 'spl
c', 'says.', 'members', 'chanted', '"reclaim', 'america"', 'rallies', 'coeur', "d'al
ene,", 'idaho,', 'washington', 'boston', 'recent', 'years,', 'reports.', 'patriot',
'front', 'responsible', '"the', 'vast', 'majority', 'white', 'supremacist', 'propaga
nda', 'distributed', 'united', 'states"', 'since', '2019,', 'anti', 'defamation', 'l
eague.', 'first', 'time', 'language', 'promoting', 'patriot', 'front', 'made', 'wa
y', 'public', 'space', 'alabama.', 'july,', 'graffiti', 'beneath', 'birmingham', 'br
idge', 'appeared', '"patriot', 'front', 'us"', 'spray', 'painted', 'red', 'blue', 'l
etters,', 'al.com', 'reported.', 'patriot', 'front', 'graffiti', 'also', 'spotted',
'birmingham,', 'city', 'population', "that's", 'nearly', '70', 'percent', 'black,',

```
'u.s.', 'census', 'data.', 'photo', 'posted', 'twitter', 'month', 'showed', 'patrio
t', 'front', 'graffiti', 'along', 'red', 'mountain', 'expressway', 'birmingham', 'wo
rds,', '"we', 'dare', 'defend', 'rights."', 'patriot', 'front', 'graffiti', 'later',
'removed,', 'message', 'left', 'sydney', 'duncan,', 'attorney', 'director', 'magic',
'legal', 'center', 'birmingham,', 'saddened', 'hate', 'become', 'public', 'parts',
'alabama.', '"white', 'supremacy', 'alive', 'well,"', 'duncan', 'wrote.', 'hughes',
'traveling', 'north', 'birmingham', 'pulled', '65', 'take', 'photos', 'messages', 's
ign.', 'seen', 'confederate', 'monuments', 'flags', 'drive', 'before,', 'kind', 'mes
saging', 'government', 'owned', 'property', 'different,', 'said.', 'police', 'office
r', 'already', 'scene', 'waved', 'keep', 'driving,', 'hughes', 'added.', 'returned',
'home,', 'hughes', 'felt', 'compelled', 'share', 'images', 'due', 'ongoing', 'conver
sation', 'happening', 'among', 'birmingham', 'residents', 'promotion', 'patriot', 'f
ront', 'public', 'spaces.', '"some', 'people', 'might', 'perceive', 'upsetting', 'sc
ary,', 'sign', 'worsening', 'country,"', 'said.', '"but', 'strategy,', 'really', 'im
pressed."', 'added,', '"they\'re', 'dying', 'breed."', 'toluse', 'olorunnipa', 'az
```

Display no stop words

In [36]:
```
#uniq_tok(df_col=slct_tbl_full_df03['no_sw'])
```

Rejoin semi_processed tokens

In [37]:
```
slct_tbl_full_df03['no_sw_join'] = slct_tbl_full_df03['no_sw'].apply(" ".join)

print(slct_tbl_full_df03.shape)
display(slct_tbl_full_df03.head())

for c in range(0,1):
    print(slct_tbl_full_df03['no_sw_join'][c])
```

```
(4026, 19)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

travelers alabama driving interstate 65 parties barbecues memorial day might seen me ssages digital road signs honoring veterans died fighting united states. that's driv ers near clanton, ala., saw monday. instead, motorists seeing sign apparently hacked display words "reclaim america," white nationalist slogan, "patriot front us," refer encing white supremacist group involved deadly 2017 unite right rally charlottesvill e. "how come about?" wrote sarah hughes, motorist captured photos sign posted twitte r. "weird hell." contractor's portable message board hacked 65 chilton county, ala., monday afternoon, john mcwilliams, spokesman alabama department transportation (aldo t) west central region, washington post statement. "a citizen alerted nearby state t rooper message, contacted aldot," mcwilliams tuesday. "aldot personnel immediately r esponded turned message board off. message boards 65 affected." mcwilliams added ald ot investigating white supremacist language appeared sign near clanton, 40 miles nor thwest montgomery, ala. officials given immediate indication responsible apparently hacking interstate sign. first al.com. hughes post driving home birmingham weekend a labama's gulf coast saw white supremacist messages recently popped around home city supporters patriot front. "when saw it, thought, 'oh, guys,' " hughes, 31 year old a ttorney. "i kind shocked." hacked alabama road sign comes time president biden decla red white supremacy "the dangerous terrorist threat" country. commencement address h oward university month, biden graduating class historically black university pledged "to stand poison white supremacy, inaugural address — single dangerous terrorist thr eat homeland white supremacy." "i tell progress toward justice often meets ferocious pushback oldest sinister forces," biden may 13 address, quoting donald trump's equiv ocating response 2017 rally charlottesville killed 32 year old heather heyer injured 19 others. "that's hate never goes away." biden calls white supremacy greatest terro rism threat 2024 race heats southern poverty law center (splc) tracked least 13 hate groups alabama 2021, including proud boys. discussion surrounding white supremacists white nationalists alabama intensified month sen. tommy tuberville (r ala.) people i dentified "white extremists" white nationalists allowed serve u.s. armed forces. ask ed reporter wbhm birmingham whether white nationalists allowed serve military, tuber ville replied, "well, call that. call americans." tuberville criticized, spokesman p ost senator "resents implication people military anything patriots heroes." gop sena tor says white nationalists military, 'i call americans' patriot front, white suprem acist group whose name displayed interstate sign, texas based hate group broke vangu ard america formed charlottesville rally, splc says. members chanted "reclaim americ a" rallies coeur d'alene, idaho, washington boston recent years, reports. patriot fr ont responsible "the vast majority white supremacist propaganda distributed united s tates" since 2019, anti defamation league. first time language promoting patriot fro nt made way public space alabama. july, graffiti beneath birmingham bridge appeared "patriot front us" spray painted red blue letters, al.com reported. patriot front gr affiti also spotted birmingham, city population that's nearly 70 percent black, u.s. census data. photo posted twitter month showed patriot front graffiti along red moun tain expressway birmingham words, "we dare defend rights." patriot front graffiti la ter removed, message left sydney duncan, attorney director magic legal center birmin gham, saddened hate become public parts alabama. "white supremacy alive well," dunca n wrote. hughes traveling north birmingham pulled 65 take photos messages sign. seen confederate monuments flags drive before, kind messaging government owned property d ifferent, said. police officer already scene waved keep driving, hughes added. retur ned home, hughes felt compelled share images due ongoing conversation happening amon g birmingham residents promotion patriot front public spaces. "some people might per ceive upsetting scary, sign worsening country," said. "but strategy, really impresse

Remove punctuation

In [38]:
```
punctuation = set(punctuation) # speeds up comparison
#print(punctuation)

# Add special hyphen mark
tw_punct = punctuation - {"#"}
#print(tw_punct)

# Remove hash and at symbols for later capture of hashtag info
tw_punct = tw_punct - {"@"}
tw_punct = tw_punct - {"-"}
#tw_punct = tw_punct - {"/"}
tw_punct.add("’")
tw_punct.add("‘")
tw_punct.add("”")
tw_punct.add("“")
tw_punct.add("…")
tw_punct.add("—")
tw_punct.add("...")
tw_punct.add("€")
tw_punct.add("±")
tw_punct.add("£")
tw_punct.add("¡")
tw_punct.add("§")
tw_punct.add("◉")

print(tw_punct)
```

```
{',', ':', '+', '~', '‘', ';', "'", '}', '.', '¡', '’', '€', '_', '/', ']', '§',
'?', '{', '◉', '±', '%', '(', '...', '!', '$', '[', '=', '>', '”', ')', '“', '`',
'&', '^', '\\', '—', '*', '£', '|', '"', '<', '…'}
```

Create function

In [39]:
```
def remove_punctuation(text, punct_set=tw_punct):
    return("".join([ch for ch in text if ch not in punct_set]))
```

Call function

In [40]:
```
slct_tbl_full_df03['no_sw_join_no_punc'] = slct_tbl_full_df03['no_sw_join']\
.apply(remove_punctuation, punct_set=tw_punct)

print(slct_tbl_full_df03.shape)
display(slct_tbl_full_df03.head())

for c in range(0,1):
    try:
        print(slct_tbl_full_df03['no_sw_join_no_punc'][c], '\n')
    except:
        print(f'\nerror on {c}\n')
```

```
(4026, 20)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

travelers alabama driving interstate 65 parties barbecues memorial day might seen me ssages digital road signs honoring veterans died fighting united states thats driver s near clanton ala saw monday instead motorists seeing sign apparently hacked displa y words reclaim america white nationalist slogan patriot front us referencing white supremacist group involved deadly 2017 unite right rally charlottesville how come ab out wrote sarah hughes motorist captured photos sign posted twitter weird hell contr actors portable message board hacked 65 chilton county ala monday afternoon john mcw illiams spokesman alabama department transportation aldot west central region washin gton post statement a citizen alerted nearby state trooper message contacted aldot m cwilliams tuesday aldot personnel immediately responded turned message board off mes sage boards 65 affected mcwilliams added aldot investigating white supremacist langu age appeared sign near clanton 40 miles northwest montgomery ala officials given imm ediate indication responsible apparently hacking interstate sign first alcom hughes post driving home birmingham weekend alabamas gulf coast saw white supremacist messa ges recently popped around home city supporters patriot front when saw it thought oh guys hughes 31 year old attorney i kind shocked hacked alabama road sign comes time president biden declared white supremacy the dangerous terrorist threat country comm encement address howard university month biden graduating class historically black u niversity pledged to stand poison white supremacy inaugural address single dangerou s terrorist threat homeland white supremacy i tell progress toward justice often mee ts ferocious pushback oldest sinister forces biden may 13 address quoting donald tru mps equivocating response 2017 rally charlottesville killed 32 year old heather heye r injured 19 others thats hate never goes away biden calls white supremacy greatest terrorism threat 2024 race heats southern poverty law center splc tracked least 13 h ate groups alabama 2021 including proud boys discussion surrounding white supremacis ts white nationalists alabama intensified month sen tommy tuberville r ala people id entified white extremists white nationalists allowed serve us armed forces asked rep orter wbhm birmingham whether white nationalists allowed serve military tuberville r eplied well call that call americans tuberville criticized spokesman post senator re sents implication people military anything patriots heroes gop senator says white na tionalists military i call americans patriot front white supremacist group whose nam e displayed interstate sign texas based hate group broke vanguard america formed cha rlottesville rally splc says members chanted reclaim america rallies coeur dalene id aho washington boston recent years reports patriot front responsible the vast majori ty white supremacist propaganda distributed united states since 2019 anti defamation league first time language promoting patriot front made way public space alabama jul y graffiti beneath birmingham bridge appeared patriot front us spray painted red blu e letters alcom reported patriot front graffiti also spotted birmingham city populat ion thats nearly 70 percent black us census data photo posted twitter month showed p atriot front graffiti along red mountain expressway birmingham words we dare defend rights patriot front graffiti later removed message left sydney duncan attorney dire ctor magic legal center birmingham saddened hate become public parts alabama white s upremacy alive well duncan wrote hughes traveling north birmingham pulled 65 take ph otos messages sign seen confederate monuments flags drive before kind messaging gove rnment owned property different said police officer already scene waved keep driving hughes added returned home hughes felt compelled share images due ongoing conversati on happening among birmingham residents promotion patriot front public spaces some p eople might perceive upsetting scary sign worsening country said but strategy really impressed added theyre dying breed toluse olorunnipa azi paybarah contributed report

Tokenize

In [41]:
```
slct_tbl_full_df03['no_sw_join_no_punc_tok'] \
= slct_tbl_full_df03['no_sw_join_no_punc'].apply(str.split)
```

```
print(slct_tbl_full_df03.shape)
display(slct_tbl_full_df03.head())

for c in range(0,1):
    print(slct_tbl_full_df03['no_sw_join_no_punc_tok'][c], '\n')
```

```
(4026, 21)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

5 rows × 21 columns

['travelers', 'alabama', 'driving', 'interstate', '65', 'parties', 'barbecues', 'mem
orial', 'day', 'might', 'seen', 'messages', 'digital', 'road', 'signs', 'honoring',
'veterans', 'died', 'fighting', 'united', 'states', 'thats', 'drivers', 'near', 'cla
nton', 'ala', 'saw', 'monday', 'instead', 'motorists', 'seeing', 'sign', 'apparentl
y', 'hacked', 'display', 'words', 'reclaim', 'america', 'white', 'nationalist', 'slo
gan', 'patriot', 'front', 'us', 'referencing', 'white', 'supremacist', 'group', 'inv
olved', 'deadly', '2017', 'unite', 'right', 'rally', 'charlottesville', 'how', 'com
e', 'about', 'wrote', 'sarah', 'hughes', 'motorist', 'captured', 'photos', 'sign',
'posted', 'twitter', 'weird', 'hell', 'contractors', 'portable', 'message', 'board',
'hacked', '65', 'chilton', 'county', 'ala', 'monday', 'afternoon', 'john', 'mcwillia
ms', 'spokesman', 'alabama', 'department', 'transportation', 'aldot', 'west', 'centr
al', 'region', 'washington', 'post', 'statement', 'a', 'citizen', 'alerted', 'nearb
y', 'state', 'trooper', 'message', 'contacted', 'aldot', 'mcwilliams', 'tuesday', 'a
ldot', 'personnel', 'immediately', 'responded', 'turned', 'message', 'board', 'off',
'message', 'boards', '65', 'affected', 'mcwilliams', 'added', 'aldot', 'investigatin
g', 'white', 'supremacist', 'language', 'appeared', 'sign', 'near', 'clanton', '40',
'miles', 'northwest', 'montgomery', 'ala', 'officials', 'given', 'immediate', 'indic
ation', 'responsible', 'apparently', 'hacking', 'interstate', 'sign', 'first', 'alco
m', 'hughes', 'post', 'driving', 'home', 'birmingham', 'weekend', 'alabamas', 'gul
f', 'coast', 'saw', 'white', 'supremacist', 'messages', 'recently', 'popped', 'aroun
d', 'home', 'city', 'supporters', 'patriot', 'front', 'when', 'saw', 'it', 'though
t', 'oh', 'guys', 'hughes', '31', 'year', 'old', 'attorney', 'i', 'kind', 'shocked',
'hacked', 'alabama', 'road', 'sign', 'comes', 'time', 'president', 'biden', 'declare
d', 'white', 'supremacy', 'the', 'dangerous', 'terrorist', 'threat', 'country', 'com
mencement', 'address', 'howard', 'university', 'month', 'biden', 'graduating', 'clas
s', 'historically', 'black', 'university', 'pledged', 'to', 'stand', 'poison', 'whit
e', 'supremacy', 'inaugural', 'address', 'single', 'dangerous', 'terrorist', 'threa
t', 'homeland', 'white', 'supremacy', 'i', 'tell', 'progress', 'toward', 'justice',
'often', 'meets', 'ferocious', 'pushback', 'oldest', 'sinister', 'forces', 'biden',
'may', '13', 'address', 'quoting', 'donald', 'trumps', 'equivocating', 'response',
'2017', 'rally', 'charlottesville', 'killed', '32', 'year', 'old', 'heather', 'heye
r', 'injured', '19', 'others', 'thats', 'hate', 'never', 'goes', 'away', 'biden', 'c
alls', 'white', 'supremacy', 'greatest', 'terrorism', 'threat', '2024', 'race', 'hea
ts', 'southern', 'poverty', 'law', 'center', 'splc', 'tracked', 'least', '13', 'hat
e', 'groups', 'alabama', '2021', 'including', 'proud', 'boys', 'discussion', 'surrou
nding', 'white', 'supremacists', 'white', 'nationalists', 'alabama', 'intensified',
'month', 'sen', 'tommy', 'tuberville', 'r', 'ala', 'people', 'identified', 'white',
'extremists', 'white', 'nationalists', 'allowed', 'serve', 'us', 'armed', 'forces',
'asked', 'reporter', 'wbhm', 'birmingham', 'whether', 'white', 'nationalists', 'allo
wed', 'serve', 'military', 'tuberville', 'replied', 'well', 'call', 'that', 'call',
'americans', 'tuberville', 'criticized', 'spokesman', 'post', 'senator', 'resents',
'implication', 'people', 'military', 'anything', 'patriots', 'heroes', 'gop', 'senat
or', 'says', 'white', 'nationalists', 'military', 'i', 'call', 'americans', 'patrio
t', 'front', 'white', 'supremacist', 'group', 'whose', 'name', 'displayed', 'interst
ate', 'sign', 'texas', 'based', 'hate', 'group', 'broke', 'vanguard', 'america', 'fo
rmed', 'charlottesville', 'rally', 'splc', 'says', 'members', 'chanted', 'reclaim',
'america', 'rallies', 'coeur', 'dalene', 'idaho', 'washington', 'boston', 'recent',
'years', 'reports', 'patriot', 'front', 'responsible', 'the', 'vast', 'majority', 'w
hite', 'supremacist', 'propaganda', 'distributed', 'united', 'states', 'since', '201
9', 'anti', 'defamation', 'league', 'first', 'time', 'language', 'promoting', 'patri
ot', 'front', 'made', 'way', 'public', 'space', 'alabama', 'july', 'graffiti', 'bene
ath', 'birmingham', 'bridge', 'appeared', 'patriot', 'front', 'us', 'spray', 'painte
d', 'red', 'blue', 'letters', 'alcom', 'reported', 'patriot', 'front', 'graffiti',
'also', 'spotted', 'birmingham', 'city', 'population', 'thats', 'nearly', '70', 'per
cent', 'black', 'us', 'census', 'data', 'photo', 'posted', 'twitter', 'month', 'show
ed', 'patriot', 'front', 'graffiti', 'along', 'red', 'mountain', 'expressway', 'birm

```
ingham', 'words', 'we', 'dare', 'defend', 'rights', 'patriot', 'front', 'graffiti',
'later', 'removed', 'message', 'left', 'sydney', 'duncan', 'attorney', 'director',
'magic', 'legal', 'center', 'birmingham', 'saddened', 'hate', 'become', 'public', 'p
arts', 'alabama', 'white', 'supremacy', 'alive', 'well', 'duncan', 'wrote', 'hughe
s', 'traveling', 'north', 'birmingham', 'pulled', '65', 'take', 'photos', 'message
s', 'sign', 'seen', 'confederate', 'monuments', 'flags', 'drive', 'before', 'kind',
'messaging', 'government', 'owned', 'property', 'different', 'said', 'police', 'offi
cer', 'already', 'scene', 'waved', 'keep', 'driving', 'hughes', 'added', 'returned',
'home', 'hughes', 'felt', 'compelled', 'share', 'images', 'due', 'ongoing', 'convers
ation', 'happening', 'among', 'birmingham', 'residents', 'promotion', 'patriot', 'fr
ont', 'public', 'spaces', 'some', 'people', 'might', 'perceive', 'upsetting', 'scar
y', 'sign', 'worsening', 'country', 'said', 'but', 'strategy', 'really', 'impresse
d', 'added', 'theyre', 'dying', 'breed', 'toluse', 'olorunnipa', 'azi', 'paybarah',
'contributed', 'report']
```

Display globally unique tokens on final tokens

In [42]:
```
#uniq_tok(df_col=slct_tbl_full_df03['no_sw_join_no_punc_tok'])
```

Pipeline consolidation

Pipeline function

In [43]:
```
def prepare(text, pipeline):
    '''Run a pipeline of text processing transformers'''
    tokens = str(text)

    # Pull key and val from trans dictionaries
    for transformer in pipeline:
        trans = list(transformer.keys())[0]
        args = list(transformer.values())[0]
        #print(trans)
        #print(args)
        if args == None:
            #print(1)
            tokens = trans(tokens)
        else:
            #print('check99', trans, args)
            tokens = trans(tokens, args)

    return(tokens)
```

article_text preprocessing _ w/o lemmatization

In [44]:
```
'''Set transformer pipeline 1:
Caseloading, normalization (using textacy), special ch removal,
split on whitespace, stop word removal, rejoin,
remove custom punctuation, tokenize
'''
transformers01 = [{str.lower: None},
                  {normalize: None},
```

```
                    {rex_replace: None},
                    {rex_url: None},
                    {emoji_split: None},
                    {str.split: None},
                    {sw_remover: None},
                    {" ".join: None},
                    {remove_punctuation: tw_punct},
                    {str.split: None},
                    {" ".join: None},
                   ]

# Apply transformers to pandas dataframe, w/ new col containing tokens
slct_tbl_full_df04['processed_text'] = slct_tbl_full_df04['article_text']\
.progress_apply(prepare, pipeline=transformers01)

slct_tbl_full_df04['processed_text_split'] = slct_tbl_full_df04['processed_text']\
.progress_apply(str.split)

slct_tbl_full_df04['num_tokens'] = slct_tbl_full_df04['processed_text_split']\
.map(len)

display(slct_tbl_full_df04.head())

# Review unique tokens across entire dataset
for c in range(0,1):
    try:
        print(slct_tbl_full_df04['processed_text'][c], '\n')
    except:
        print(f'Skip {c}')
```

```
100%|████████| 4026/4026 [00:27<00:00, 143.88it/s]
100%|██████| 4026/4026 [00:00<00:00, 15319.36it/s]
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

travelers alabama driving interstate 65 parties barbecues memorial day might seen me ssages digital road signs honoring veterans died fighting united states thats driver s near clanton ala saw monday instead motorists seeing sign apparently hacked displa y words reclaim america white nationalist slogan patriot front us referencing white supremacist group involved deadly 2017 unite right rally charlottesville how come ab out wrote sarah hughes motorist captured photos sign posted twitter weird hell contr actors portable message board hacked 65 chilton county ala monday afternoon john mcw illiams spokesman alabama department transportation aldot west central region washin gton post statement a citizen alerted nearby state trooper message contacted aldot m cwilliams tuesday aldot personnel immediately responded turned message board off mes sage boards 65 affected mcwilliams added aldot investigating white supremacist langu age appeared sign near clanton 40 miles northwest montgomery ala officials given imm ediate indication responsible apparently hacking interstate sign first alcom hughes post driving home birmingham weekend alabamas gulf coast saw white supremacist messa ges recently popped around home city supporters patriot front when saw it thought oh guys hughes 31 year old attorney i kind shocked hacked alabama road sign comes time president biden declared white supremacy the dangerous terrorist threat country comm encement address howard university month biden graduating class historically black u niversity pledged to stand poison white supremacy inaugural address single dangerous terrorist threat homeland white supremacy i tell progress toward justice often meets ferocious pushback oldest sinister forces biden may 13 address quoting donald trumps equivocating response 2017 rally charlottesville killed 32 year old heather heyer in jured 19 others thats hate never goes away biden calls white supremacy greatest terr orism threat 2024 race heats southern poverty law center splc tracked least 13 hate groups alabama 2021 including proud boys discussion surrounding white supremacists w hite nationalists alabama intensified month sen tommy tuberville r ala people identi fied white extremists white nationalists allowed serve us armed forces asked reporte r wbhm birmingham whether white nationalists allowed serve military tuberville repli ed well call that call americans tuberville criticized spokesman post senator resent s implication people military anything patriots heroes gop senator says white nation alists military i call americans patriot front white supremacist group whose name di splayed interstate sign texas based hate group broke vanguard america formed charlot tesville rally splc says members chanted reclaim america rallies coeur dalene idaho washington boston recent years reports patriot front responsible the vast majority w hite supremacist propaganda distributed united states since 2019 anti defamation lea gue first time language promoting patriot front made way public space alabama july g raffiti beneath birmingham bridge appeared patriot front us spray painted red blue l etters alcom reported patriot front graffiti also spotted birmingham city population thats nearly 70 percent black us census data photo posted twitter month showed patri ot front graffiti along red mountain expressway birmingham words we dare defend righ ts patriot front graffiti later removed message left sydney duncan attorney director magic legal center birmingham saddened hate become public parts alabama white suprem acy alive well duncan wrote hughes traveling north birmingham pulled 65 take photos messages sign seen confederate monuments flags drive before kind messaging governmen t owned property different said police officer already scene waved keep driving hugh es added returned home hughes felt compelled share images due ongoing conversation h appening among birmingham residents promotion patriot front public spaces some peopl e might perceive upsetting scary sign worsening country said but strategy really imp ressed added theyre dying breed toluse olorunnipa azi paybarah contributed report

Display globally unqiue tokens on final tokens

In [45]:
```
#uniq_tok(df_col=slct_tbl_full_df04['processed_text_split'])
```

article_text preprocessing_w/ lemmatization

'''Set transformer pipeline 2: Caseloading, normalization (using textacy), special ch removal, lemmitization, stop word removal, rejoin, remove custom punctuation, tokenize ''' transformers02 = [{str.lower: None}, {normalize: None}, {rex_replace: None}, {lemma: None}, {" ".join: None}, {rex_url: None}, {emoji_split: None}, {str.split: None}, {sw_remover: None}, {" ".join: None}, {remove_punctuation: tw_punct}, {str.split: None}, {" ".join: None},] # Apply transformers to pandas dataframe, w/ new col containing tokens slct_tbl_full_df04['processed_lemmas'] = slct_tbl_full_df04['article_text']\ .progress_apply(prepare, pipeline=transformers02) slct_tbl_full_df04['processed_lemmas_split'] = slct_tbl_full_df04['processed_lemmas']\ .progress_apply(str.split) slct_tbl_full_df04['num_lemmas'] = slct_tbl_full_df04['processed_lemmas_split']\ .map(len) display(slct_tbl_full_df04.head()) # Review unique tokens across entire dataset for c in range(0,1): try: print(slct_tbl_full_df04['processed_lemmas'][c], '\n') except: print(f'Skip {c}')

Display globally unqiue tokens on final tokens

In [46]:
```
#uniq_tok(df_col=slct_tbl_full_df04['processed_lemmas_split'])
```

Calculate concentration ratio of each set of corpora

In [47]:
```
display(slct_tbl_full_df04['political_lean'].value_counts())

slct_tbl_full_df04_left = slct_tbl_full_df04[\
                                            slct_tbl_full_df04[\
                                               'political_lean'] == 'left']

print(slct_tbl_full_df04_left.shape)
#display(slct_tbl_full_df04_left.head())

slct_tbl_full_df04_right = slct_tbl_full_df04[\
                                             slct_tbl_full_df04[\
                                               'political_lean'] == 'right']

print(slct_tbl_full_df04_right.shape)
#display(slct_tbl_full_df04_right.head())

slct_tbl_full_df04_left_s1 = list(itertools.chain.from_iterable(
    list(pd.Series(slct_tbl_full_df04_left['processed_text_split']))))
print(slct_tbl_full_df04_left_s1[:10])
slct_tbl_full_df04_right_s1 = list(itertools.chain.from_iterable(
    list(pd.Series(slct_tbl_full_df04_right['processed_text_split']))))
print(slct_tbl_full_df04_right_s1[:10])
```

```
right    2758
left     1268
Name: political_lean, dtype: int64
(1268, 14)
(2758, 14)
['travelers', 'alabama', 'driving', 'interstate', '65', 'parties', 'barbecues', 'mem
orial', 'day', 'might']
['family', 'jennifer', 'farber', 'dulos', 'released', 'statement', 'wednesday', 'mar
king', 'four', 'years']
```

In [48]:
```
def concen_ratio(artist_lst=[],
                 lsts=[]):
```

```
lyr_corp_lst = []
for l in lsts:
    print(type(l))
    lyr_corp_lst.append(' '.join(l))
print(len(lyr_corp_lst))
#print(lyr_corp_lst)

cv = CountVectorizer(input='content',
                     encoding='utf-8',
                     stop_words=None,
                     token_pattern=r'\S+'
                    )

lyr_tokens_fit = cv.fit(lyr_corp_lst)

print(pd.Series(cv.get_feature_names_out()).sample(15))

lyr_tokens_sm = cv.transform(lyr_corp_lst)
display(lyr_tokens_sm)

df = pd.DataFrame(lyr_tokens_sm.toarray(),
                  columns=cv.get_feature_names_out())
#display(df)

df02 = df.copy()
df02['r_sum'] = df02.sum(axis=1)
#display(df02)

'''Filter by frequency for all columns citation:
OpenAI. (2021). ChatGPT [Computer software]. https://openai.com/;
https://pandas.pydata.org/pandas-docs/stable/reference/api/
pandas.DataFrame.ge.html'''
condition = df.ge(5).all()
#print(condition)

# Get the list of columns that satisfy the condition
columns = condition[condition].index.tolist()
#print(columns)
columns.append('r_sum')
#print(columns)

#display(df02[columns])

df03 = df02[columns].copy()
display(df03)

'''Filter by frequency for all columns & add summary row citation:
OpenAI. (2021). ChatGPT [Computer software]. https://openai.com/'''
df04 = df03.apply(lambda x: x / df03.iloc[:,-1], axis=0)
#display(df04)

# Create new rows by dividing one artist row by the second artists row
new_row01 = df04.iloc[0] / df04.iloc[1]
new_row02 = df04.iloc[1] / df04.iloc[0]

# Append the new row to the DataFrame
```

```
    df04 = df04.append(new_row01, ignore_index=True)
    df04 = df04.append(new_row02, ignore_index=True)
    display(df04)

    # Transpose dataframe
    df05 = df04.T
    df05 = df05.reset_index()
    df05.columns = ['token',
                    'c1_concen',
                    'c2_concen',
                    'c1c2_concen_ratio',
                    'c2c1_concen_ratio']
    #print(df05)

    '''Sort values citation:
    https://pandas.pydata.org/pandas-docs/stable/reference/api
    /pandas.DataFrame.sort_values.html'''
    print(artist_lst[0])
    display(df05[['token',
                  'c1c2_concen_ratio']].sort_values(by='c1c2_concen_ratio',
                                                    ascending=False).head(10))
    print(artist_lst[1])
    display(df05[['token',
                  'c2c1_concen_ratio']].sort_values(by='c2c1_concen_ratio',
                                                    ascending=False).head(10))

concen_ratio(artist_lst=['Left-Right Concentration Ratio',
                         'Right-Left Concentration Ratio'],
             lsts=[slct_tbl_full_df04_left_s1,
                   slct_tbl_full_df04_right_s1])
```

```
<class 'list'>
<class 'list'>
2
19846            feuding
20622            forgave
1809              516000
13151            coulter
6846             barroom
45411          shoplifted
48509         suspecttime
54602               wout
4016              agarro
25526             impart
3902          aeronautics
48385         supportable
14863             delores
5273           appetizing
2580     @desantiswarroom
dtype: object
<2x55704 sparse matrix of type '<class 'numpy.int64'>'
        with 79403 stored elements in Compressed Sparse Row format>
```

	#2	#metoo	0	07	1	10	100	1000	10000	100000	...	zero	zip	zone	zones	
0	6	18	16	5	452	397	138	65	52	45	...	67	6	38	7	
1	5	8	36	6	600	638	226	82	124	64	...	97	11	38	11	

2 rows × 10958 columns

```
C:\Users\acarr\AppData\Local\Temp\ipykernel_23812\3142308763.py:59: FutureWarning: T
he frame.append method is deprecated and will be removed from pandas in a future ver
sion. Use pandas.concat instead.
  df04 = df04.append(new_row01, ignore_index=True)
C:\Users\acarr\AppData\Local\Temp\ipykernel_23812\3142308763.py:60: FutureWarning: T
he frame.append method is deprecated and will be removed from pandas in a future ver
sion. Use pandas.concat instead.
  df04 = df04.append(new_row02, ignore_index=True)
```

	#2	#metoo	0	07	1	10	100	1000	
0	7.9407e-06	2.3822e-05	2.1175e-05	6.6172e-06	0.0006	0.0005	0.0002	8.6024e-05	6.88
1	4.7559e-06	7.6094e-06	3.4242e-05	5.7070e-06	0.0006	0.0006	0.0002	7.7996e-05	1.17
2	1.6697e+00	3.1306e+00	6.1839e-01	1.1595e+00	1.0482	0.8658	0.8496	1.1029e+00	5.83
3	5.9893e-01	3.1943e-01	1.6171e+00	8.6245e-01	0.9540	1.1550	1.1770	9.0668e-01	1.713

4 rows × 10958 columns

Left-Right Concentration Ratio

	token	c1c2_concen_ratio
723	anonymity	17.3923
9885	thomass	16.1400
10774	willis	14.1921
6201	mehta	13.6046
3115	docket	13.0790
4349	ginni	12.2442
7421	pork	11.5948
10224	uncertainty	9.9537
4873	hush	9.7397
4400	gorsuch	9.7397

Right-Left Concentration Ratio

	token	c2c1_concen_ratio
755	ap	19.4052
6725	nyc	17.4886
8200	reparations	16.8897
579	aliens	12.6972
1699	ccp	11.4994
5910	locker	11.0681
2057	comey	9.5828
5975	lula	9.1995
1530	busch	8.8940
3957	fidelity	8.7683

KWIC

In [49]:
```
def kwic(doc_series, keyword, window=35, print_samples=5):
    '''Search article text for keywords in context (KWIC)'''
    def add_kwic(text):
        kwic_list.extend(keyword_in_context(doc=text,
                                            keyword=keyword,
                                            ignore_case=True,
                                            window_width=window))
    kwic_list = []
    doc_series.map(add_kwic)

    if print_samples is None or print_samples==0:
        return kwic_list
    else:
        k = min(print_samples, len(kwic_list))
        print(f"{k} random samples out of {len(kwic_list)}" + \
              f" contexts for '{keyword}':")
        for sample in random.sample(list(kwic_list), k):
            print(re.sub(r'[\n\t]', ' ', sample[0]) + ' ' +\
                  sample[1] + ' ' +\
                  re.sub(r'[\n\t]', ' ', sample[2]))
```

In [50]:
```
#kwic(slct_tbl_full_df03['article_text'], 'amp', window=150, print_samples=120)

kwic(slct_tbl_full_df03['norm'], 'economic', window=50, print_samples=20)
```

```
20 random samples out of 947 contexts for 'economic':
government took one step back from self-inflicted  economic  disaster on tuesday. ho
use republicans avoided th
re black americans have significant political and  economic  clout. the resulting
displace
rtainty to data flows that underpin transatlantic  economic  ties, society, and our
international cooperation.
egon have experienced homelessness as a result of  economic  hardship, a shortage of
safe and affordable housi
while, trump has long polled better than biden on  economic  issues and immigration,
with americans more confi
orhood. the operation was eventually called off.   economic  crisis   the cricket le
gend-turned-politician has
ou. that's also why they've cut off access to key  economic  databases and cracked d
own on journalists. click
 facts are clear. almost every element of biden's  economic  policy has a "buy ameri
ca" component to it. its g
hold their labor is protected by the nlra even if  economic  injury results." althou
gh barrett's opinion occup
an, who is battling for a third term, buffeted by  economic  headwinds and criticism
that the impact of the fe
 areas: health care, health behaviors, social and  economic  factors, environment an
d public policy. "we hoped
we flatter condemn us to ostracism, to financial,  economic , and monetary weakness
forever," maduro
n criticized president biden on environmental and  economic  policies, has said he w
ould announce by year's en
tions to arrive at a bill that ultimately avoided  economic  disaster. through it al
l, some democrats have gru
                the two leaders called for closer  economic  and security cooperatio
n amid growing threats fro
ia.  the elections also take place amid a serious  economic  crisis and what analyst
s say is democratic erosio
f jobs.   "if they fail to do it, we will have an  economic  and financial catastrop
he that will be of our own
x news app she said she is also worried about the  economic  effects of these purcha
ses, especially in her hom
his deal, we now have a clear runway to sell that  economic  vision and implement hi
storic investments across
orkers, said elise gould, senior economist at the  economic  policy institute, a thi
nk tank that advocates for
```

Train/test split

In [51]:
```
slct_tbl_full_df04['stratifier'] = slct_tbl_full_df04['political_lean']\
.astype(str) + slct_tbl_full_df04['source_name'].astype(str)
slct_tbl_full_df04['stratifier'] = slct_tbl_full_df04['stratifier']\
.map(str.lower)
display(slct_tbl_full_df04.head())

y01a = ['stratifier']
slct_tbl_full_df04_y01_vc01a = slct_tbl_full_df04[y01a].to_numpy()
print(slct_tbl_full_df04_y01_vc01a.shape)

y01 = ['political_lean']
```

```
slct_tbl_full_df04_y01_vc01 = slct_tbl_full_df04[y01].to_numpy()
print(slct_tbl_full_df04_y01_vc01.shape)
```

	index	source_name	author	title	u
0	0	The Washington Post	NaN	Alabama Highway sign hacked with white suprema...	https://www.washingtonpost.com/nation/2023/05/
1	1	The Washington Post	Amber Phillips	Breaking down the GOP investigation into the B...	https://www.washingtonpost.com/politics/2023/0
2	2	The Washington Post	David Ovalle	Appeals court paves way for Purdue Pharma opio...	https://www.washingtonpost.com/health/2023/05/
3	3	The Washington Post	Philip Bump	Trump pledges to win an immigration fight he d...	https://www.washingtonpost.com/politics/2023/0
4	5	The Washington Post	Paul Waldman	Why fear of change will drive the GOP presiden...	https://www.washingtonpost.com/opinions/2023/0

```
(4026, 1)
(4026, 1)
```

In [52]:
```
nlm_train_x01, \
nlm_test_x01, \
nlm_train_y01, \
nlm_test_y01 = train_test_split(slct_tbl_full_df04['processed_text'],
                                slct_tbl_full_df04_y01_vc01,
                                test_size=.15,
                                random_state=1699,
                                stratify=slct_tbl_full_df04_y01_vc01a
                               )

nlm_train_y01 = nlm_train_y01.ravel()
nlm_test_y01 = nlm_test_y01.ravel()

print(f'{nlm_train_x01.shape}')
print(f'{nlm_train_y01.shape}')
print(f'\n{nlm_test_x01.shape}')
print(f'{nlm_test_y01.shape}')
```

```
(3422,)
(3422,)

(604,)
(604,)
```

lem_train_x01, \ lem_test_x01, \ lem_train_y01, \ lem_test_y01 = train_test_split(slct_tbl_full_df04['processed_lemmas'], slct_tbl_full_df04_y01_vc01, test_size=.15, random_state=1699, stratify=slct_tbl_full_df04_y01_vc01a) lem_train_y01 = lem_train_y01.ravel() lem_test_y01 = lem_test_y01.ravel() print(f'{lem_train_x01.shape}') print(f'{lem_train_y01.shape}') print(f'\n{lem_test_x01.shape}') print(f'{lem_test_y01.shape}')

TF-IDF

In [53]:
```
print(slct_tbl_full_df04['processed_text'].shape)
print(slct_tbl_full_df04['processed_text'].head())
```

```
(4026,)
0    travelers alabama driving interstate 65 partie...
1    federal prosecutor may nearing decision whethe...
2    federal appeals court tuesday cleared way drug...
3    speaking orlando november 2015 republican pres...
4    look know countrys going wrong direction flori...
Name: processed_text, dtype: object
```

In [54]:
```
nlm_tfidf = TfidfVectorizer(encoding='utf-8',
                            analyzer='word',
                            stop_words=sw,
                            token_pattern=r'(?u)\b\w\w+\b',
                            ngram_range=(1,3),
                            max_df=.7,
                            min_df=5)

nlm_train_x01_mtx = nlm_tfidf.fit_transform(nlm_train_x01)
nlm_test_x01_mtx = nlm_tfidf.transform(nlm_test_x01)

display(nlm_train_x01_mtx)
display(nlm_test_x01_mtx)
```

```
<3422x48932 sparse matrix of type '<class 'numpy.float64'>'
        with 1256997 stored elements in Compressed Sparse Row format>
<604x48932 sparse matrix of type '<class 'numpy.float64'>'
        with 215091 stored elements in Compressed Sparse Row format>
```

lem_tfidf = TfidfVectorizer(encoding='utf-8', analyzer='word', stop_words=sw, token_pattern=r'(?u)\b\w\w+\b', ngram_range=(1,3), max_df=.7, min_df=5) lem_train_x01_mtx = lem_tfidf.fit_transform(lem_train_x01) lem_test_x01_mtx = lem_tfidf.transform(lem_test_x01) display(lem_train_x01_mtx) display(lem_test_x01_mtx)

In [55]:
```
def display_samp_dwm(sm=None,
                     vec=None,
                     n=(1,1),
                     rs_tup=(1,1)):
    mtx_df01 = pd.DataFrame(sm.toarray(),
                            columns=vec.get_feature_names_out())

    mtx_df01a = mtx_df01.sample(n=n[0],
                                random_state=rs_tup[0],
```

```
                                axis=1)

    mtx_df01b = mtx_df01a.sample(n=n[1],
                                 random_state=rs_tup[1],
                                 axis=0)

    display(mtx_df01b)
    return vec.get_feature_names_out()
```

In [56]:
```
rs_tup=(1699,1699)
```

In [57]:
```
nlm_train_x01_mtx_cols = display_samp_dwm(sm=nlm_train_x01_mtx,
                                          vec=nlm_tfidf,
                                          n=(17,11),
                                          rs_tup=(5,1699))
```

	nixon	follow london	efforts block	federal employees	request meeting	sexual assaults	chief executive	almost half	senior fellow	de sant
1098	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2370	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
1091	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3084	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2413	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
287	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
855	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2084	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
1537	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
991	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2245	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0

display_samp_dwm(sm=lem_train_x01_mtx, vec=lem_tfidf, n=(17,11), rs_tup=(5,1699))

Modeling

Algorithm setup

Gradient Boosting Classifier _ Using BayesSearchCV

In [ ]:
```
from sklearn.model_selection import RepeatedStratifiedKFold
```

```
# Start timer script
start_time = dt.datetime.today()

# Citation: Hochberg, 2018; Shanmukh, 2021
m2v1_gbc_pip = Pipeline([('gbc',
                          GradientBoostingClassifier(random_state=1699))])

loss_hparam = Categorical(['log_loss', 'exponential'])
lrate_hparam = Real(1e-3, 1e3, prior='log-uniform')
nest_hparam = Integer(1e2, 1e3, prior='log-uniform')
msamp_hparam = Real(.01, .95, prior='log-uniform')
mdepth_hparam = Integer(1, 20, prior='log-uniform')
mfeat_hparam = Categorical(['sqrt', 'log2', None])
#wstart_hparam = Categorical([True, False])
#calph_hparam = Real(0.0, 100.0, prior='log-uniform')
                #'gbc__warm_start': wstart_hparam
                #'gbc__ccp_alpha': calph_hparam

m2v1_gbc_grd = {'gbc__loss': loss_hparam,
                'gbc__learning_rate': lrate_hparam,
                'gbc__n_estimators': nest_hparam,
                'gbc__min_samples_split': msamp_hparam,
                'gbc__max_depth': mdepth_hparam,
                'gbc__max_features': mfeat_hparam
               }

'''Change GBC default scoring from accuracy to F1 score citation:
https://chat.openai.com/share/254f382b-4a8e-48e8-acd5-2918f0bbc59d
'''
f1_scorer = make_scorer(f1_score,
                        pos_label='right')

'''Customize cross-validation citation:
https://machinelearningmastery.com
/scikit-optimize-for-hyperparameter-tuning-in-machine-learning/
'''
cv = RepeatedStratifiedKFold(n_splits=10,
                             n_repeats=2,
                             random_state=1699)

m2v1_gbc = BayesSearchCV(m2v1_gbc_pip,
                         m2v1_gbc_grd,
                         n_iter=20,
                         scoring=f1_scorer,
                         cv=cv,
                         n_jobs=3,
                         refit=True,
                         verbose=4,
                         random_state=1699)

m2v1_gbc.fit(nlm_train_x01_mtx, nlm_train_y01)

# End timer script
end_time = dt.datetime.today()
time_elapse = end_time - start_time
print(f'Start Time = {start_time}')
```

```
print(f'End Time = {end_time}')
print(f'Elapsed Time = {time_elapse}')
```

Pickle best model

In [58]:
```
# Path to save the pickled model
mod_folder_name = 'trained_models'
m2v1_pkl_file_name = 'm2v2_gbc.pkl'

pkl_file_path01 = os.path.join(curr_dir, mod_folder_name, m2v1_pkl_file_name)

print(f'CSV file 1 in path: {pkl_file_path01}')
```

```
CSV file 1 in path: C:\Users\acarr\Documents\GitHub\ADS509_Final_project\deliverable
s\trained_models\m2v2_gbc.pkl
```

with open(pkl_file_path01, "wb") as file: pickle.dump(m2v1_gbc, file) print("Model pickled and saved successfully.")

Load pickled best model

In [59]:
```
with open(pkl_file_path01, 'rb') as file:
    m2v1_gbc = pickle.load(file)
```

In [60]:
```
print(f'\nBest Estimator:\n{m2v1_gbc.best_estimator_}')

print('\nCross-validaton results:')
display(pd.DataFrame(m2v1_gbc.cv_results_))

train_m2v1_gbc_y01_pred = m2v1_gbc.predict_proba(nlm_train_x01_mtx)
print(f'\nFirst 10 train set predictions:\n{train_m2v1_gbc_y01_pred[:10]}')

test_m2v1_gbc_y01_pred = m2v1_gbc.predict_proba(nlm_test_x01_mtx)
print(f'\nFirst 10 test set predictions:\n{test_m2v1_gbc_y01_pred[:10]}')

print(f'\nBest Score for "{m2v1_gbc.scorer_}" is {m2v1_gbc.best_score_}')
```

```
Best Estimator:
Pipeline(steps=[('gbc',
                 GradientBoostingClassifier(learning_rate=0.8373240042701702,
                                            loss='exponential', max_depth=11,
                                            max_features='sqrt',
                                            min_samples_split=0.582912491747238,
                                            random_state=1699))])

Cross-validaton results:
```

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_gbc__learning_rate
0	12.9224	1.2860	0.0274	0.0070	0.0162
1	10.5480	0.9983	0.0296	0.0091	0.035
2	2.9297	0.1453	0.0285	0.0057	55.3871
3	426.3148	43.3892	0.0356	0.0089	4.7311
4	5.6830	0.4884	0.0306	0.0039	25.6546
5	13.7448	1.6360	0.0610	0.0080	23.2484
6	11.9362	1.3784	0.0471	0.0051	173.3432
7	248.6976	19.5832	0.0346	0.0038	36.8407
8	7.2956	0.6681	0.0286	0.0028	274.3051
9	4.1264	0.3624	0.0399	0.0042	0.5714
10	4.9026	0.4603	0.0376	0.0042	0.8373
11	3.6591	0.2537	0.0421	0.0059	0.001
12	31.3106	3.3239	0.0366	0.0057	0.0012
13	531.8490	80.7165	0.0326	0.0094	0.001
14	2.5695	0.1020	0.0237	0.0081	0.0462

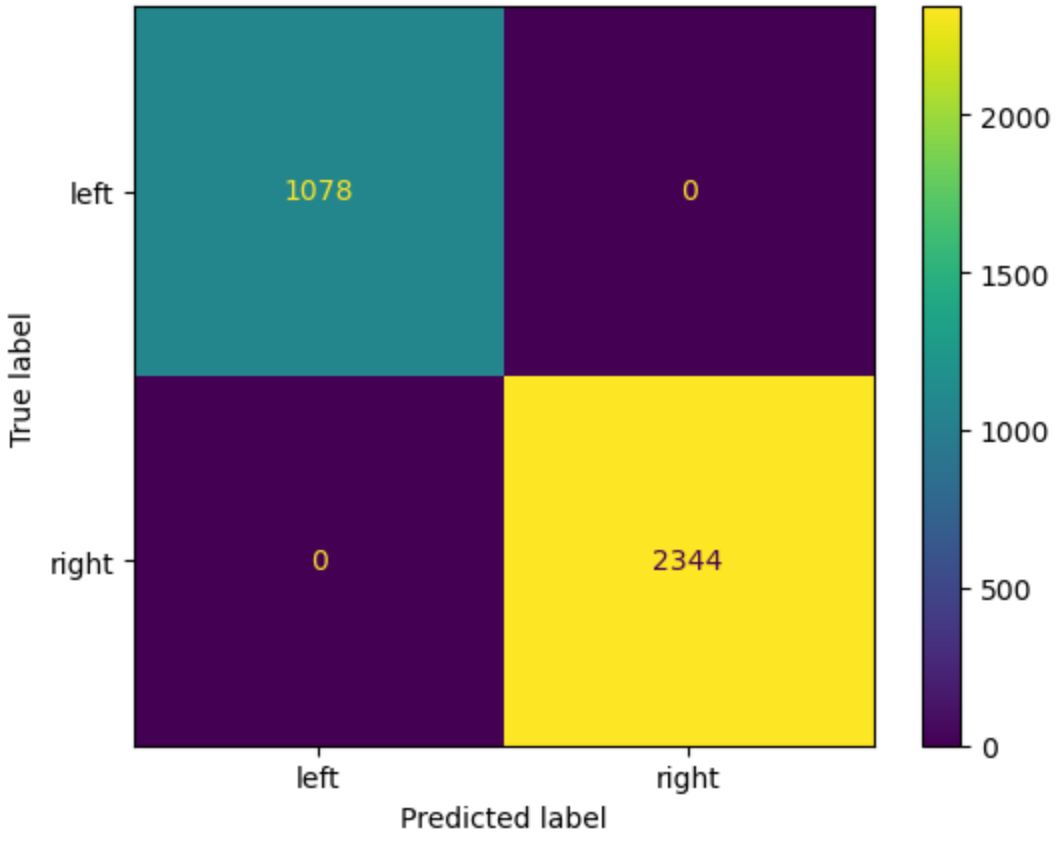
15 rows × 24 columns

```
First 10 train set predictions:
[[0.0001 0.9999]
 [0.0012 0.9988]
 [0.992  0.008 ]
 [0.     1.    ]
 [0.     1.    ]
 [0.0063 0.9937]
 [0.     1.    ]
 [0.0003 0.9997]
 [0.0023 0.9977]
 [0.9662 0.0338]]

First 10 test set predictions:
[[0.0091 0.9909]
 [0.4118 0.5882]
 [1.     0.    ]
 [0.9997 0.0003]
 [1.     0.    ]
 [0.     1.    ]
 [0.9996 0.0004]
 [0.6635 0.3365]
 [0.9917 0.0083]
 [0.0063 0.9937]]
```

Train set check

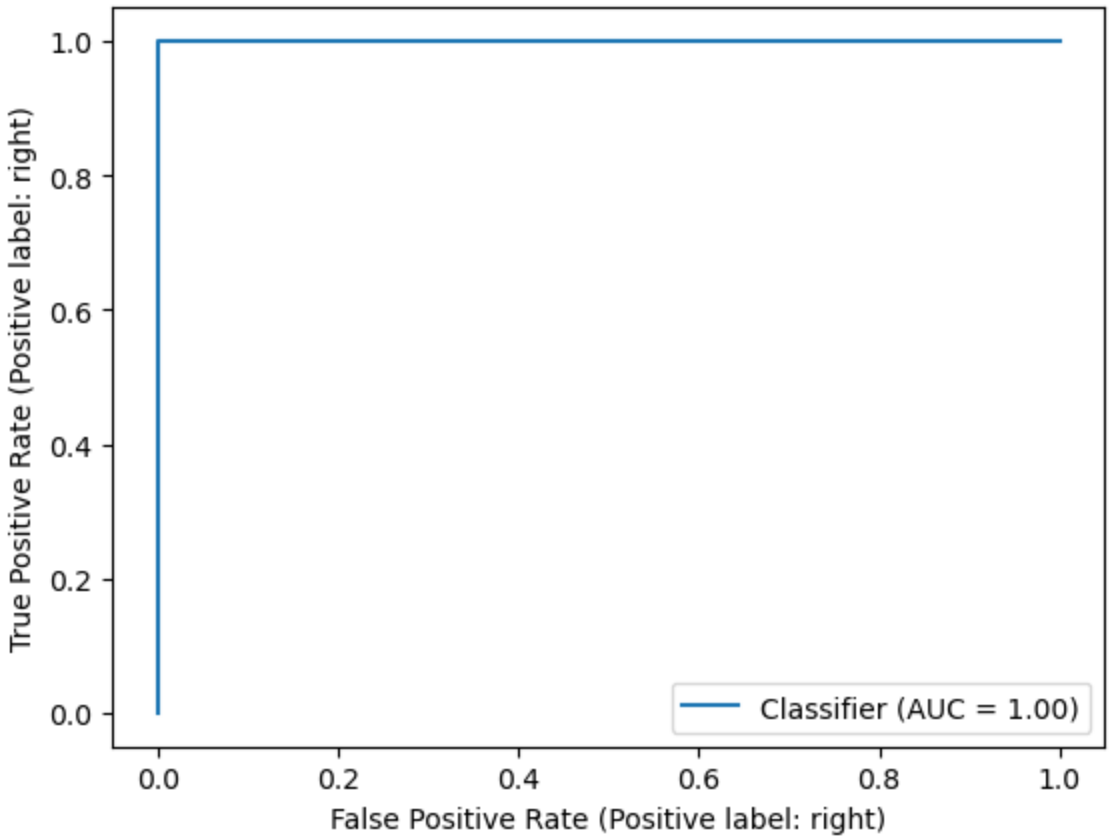
In [61]:
```
nlm_train_y01_pred = m2v1_gbc.predict(nlm_train_x01_mtx)
nlm_train_y01_pred_cm = confusion_matrix(nlm_train_y01, nlm_train_y01_pred)

print(classification_report(nlm_train_y01, nlm_train_y01_pred))
print(nlm_train_y01_pred_cm)

'''Citation:
https://scikit-learn.org/stable/modules/generated
/sklearn.metrics.ConfusionMatrixDisplay.html
#sklearn.metrics.ConfusionMatrixDisplay.plot
'''
nlm_train_cm_dsp = ConfusionMatrixDisplay(confusion_matrix=nlm_train_y01_pred_cm,
                                          display_labels=m2v1_gbc.classes_)
nlm_train_cm_dsp.plot()
plt.show()
```

```
              precision    recall  f1-score   support

        left       1.00      1.00      1.00      1078
       right       1.00      1.00      1.00      2344

    accuracy                           1.00      3422
   macro avg       1.00      1.00      1.00      3422
weighted avg       1.00      1.00      1.00      3422

[[1078    0]
 [   0 2344]]
```



ROC-AUC Curve

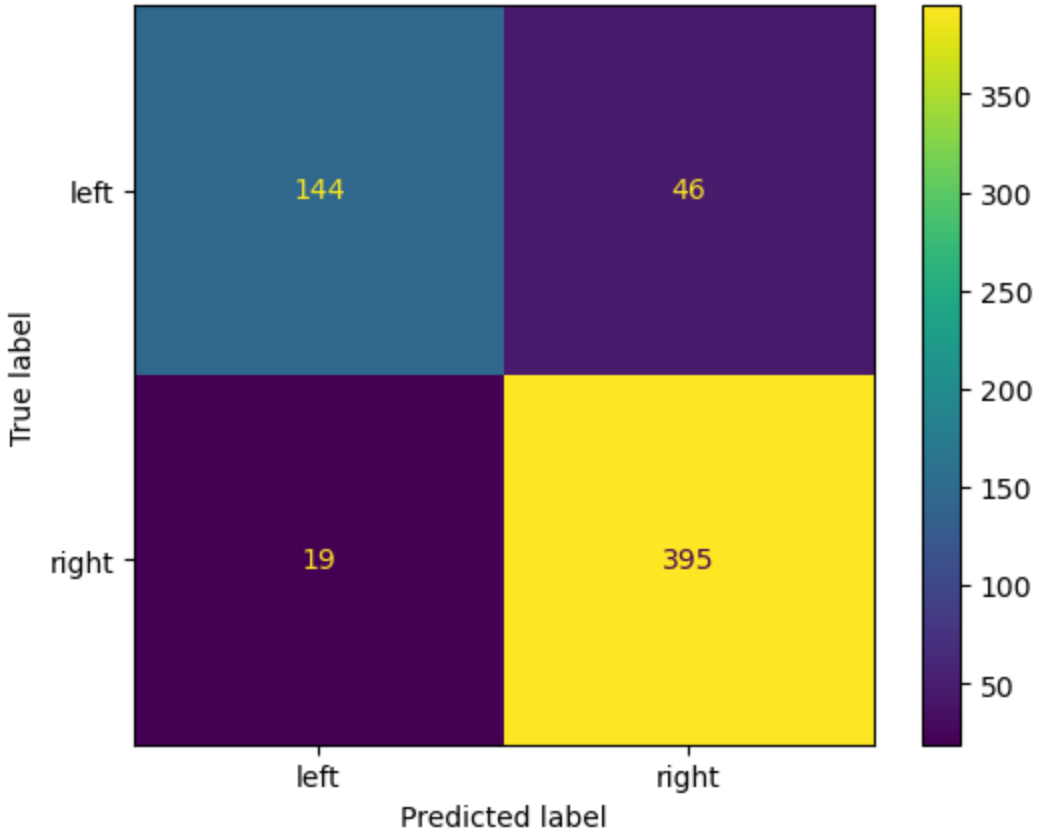
In [62]:
```
nlm_train_y01_pred_decf = m2v1_gbc.decision_function(nlm_train_x01_mtx)
RocCurveDisplay.from_predictions(nlm_train_y01, nlm_train_y01_pred_decf,
                                 pos_label='right')
plt.show()
```



Test set results

In [63]:
```
nlm_test_y01_pred = m2v1_gbc.predict(nlm_test_x01_mtx)
nlm_test_y01_pred_cm = confusion_matrix(nlm_test_y01, nlm_test_y01_pred)

print('Test Set Evaluation Metrics')
print(classification_report(nlm_test_y01, nlm_test_y01_pred))
print(nlm_test_y01_pred_cm)

'''Citation:
https://scikit-learn.org/stable/modules/generated
/sklearn.metrics.ConfusionMatrixDisplay.html
#sklearn.metrics.ConfusionMatrixDisplay.plot
'''
nlm_test_cm_dsp = ConfusionMatrixDisplay(confusion_matrix=nlm_test_y01_pred_cm,
                                         display_labels=m2v1_gbc.classes_)
nlm_test_cm_dsp.plot()
plt.show()
```

```
Test Set Evaluation Metrics
              precision    recall  f1-score   support

        left       0.88      0.76      0.82       190
       right       0.90      0.95      0.92       414

    accuracy                           0.89       604
   macro avg       0.89      0.86      0.87       604
weighted avg       0.89      0.89      0.89       604

[[144  46]
 [ 19 395]]
```



Variable importance

In [64]:

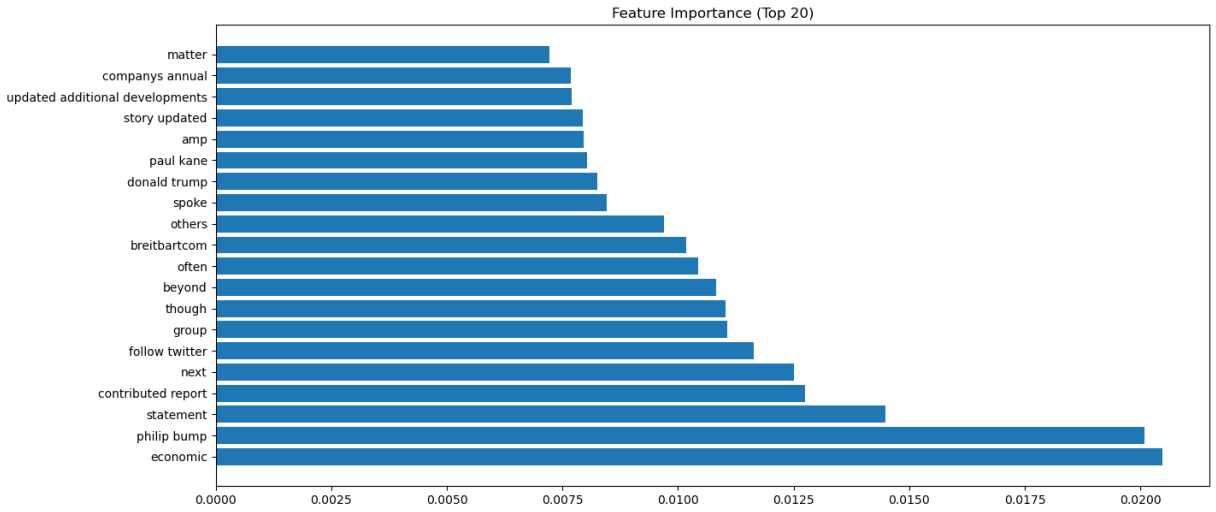
```
print(nlm_train_x01_mtx_cols)
print(type(nlm_train_x01_mtx_cols))
print(nlm_train_x01_mtx_cols.shape)

x = m2v1_gbc.best_estimator_.named_steps['gbc'].feature_importances_
x_df01 = pd.DataFrame(x, columns=['var_imp'])
x_df01['feature'] = nlm_train_x01_mtx_cols
x_df02 = x_df01.sort_values(by=['var_imp'], ascending=False)
x_df03 = x_df02.head(20)

display(x_df02.head())
print(type(x_df02))
print(x_df02.shape)
```

```
['03' '04' '05' ... 'zoos' 'zucker' 'zuckerberg']
<class 'numpy.ndarray'>
(48932,)
```

	var_imp	feature
14105	0.0205	economic
32121	0.0201	philip bump
40955	0.0145	statement
10184	0.0128	contributed report
29331	0.0125	next

```
<class 'pandas.core.frame.DataFrame'>
(48932, 2)
```

In [65]:
```
'''Citation:
https://machinelearningmastery.com/calculate-feature-importance-with-python/
'''
# plot feature importance
#figure = plt.figsize((10,9))
plt.figure(figsize=(15,7))
plt.title('Feature Importance (Top 20)')
plt.barh([x for x in range(len(x_df03['var_imp']))], x_df03['var_imp'],
         tick_label=x_df03['feature'])
plt.show()
```



In [66]:
```
TNmodel1=nlm_test_y01_pred_cm[0][0]
FPmodel1=nlm_test_y01_pred_cm[0][1]
FNmodel1=nlm_test_y01_pred_cm[1][0]
TPmodel1=nlm_test_y01_pred_cm[1][1]
```

In [67]:
```
# Results:
from tabulate import tabulate

TANmodel1=TNmodel1+FPmodel1
TAPmodel1=TPmodel1+FNmodel1
TPPmodel1=FPmodel1+TPmodel1
```

```
TPNmodel1=TNmodel1+FNmodel1
GTmodel1=TANmodel1+TAPmodel1
AccuracyM1=(TNmodel1+TPmodel1)/GTmodel1
ErrorRateM1=1-AccuracyM1
SensitivityM1=TPmodel1/(TAPmodel1)
RecallM1=SensitivityM1
SpecificityM1=TNmodel1/TANmodel1
PrecisionM1=TPmodel1/TPPmodel1
F1M1=2*PrecisionM1*RecallM1/(PrecisionM1 + RecallM1)
F2M1=5*(PrecisionM1*RecallM1)/((4*PrecisionM1)+RecallM1)
Fp5M1=(1.25)*(PrecisionM1*RecallM1)/((0.25*PrecisionM1)+RecallM1)

header = ["Accuracy", "Error Rate", "Sensitivity", "Recall", "Specificity",
          "Precision", "F1", "F2", "F0.5"]
data1 = [["Accuracy", AccuracyM1], ["Error Rate", ErrorRateM1],
         ["Sensitivity", SensitivityM1],
         ["Recall", RecallM1], ["Specificity", SpecificityM1],
         ["Precision", PrecisionM1],
         ["F1", F1M1], ["F2", F2M1], ["F0.5", Fp5M1]]

col_names=["Measurement", "Linear SVC Model"]

ModelEvaluationTable = tabulate(data1, headers=col_names,
                                tablefmt="fancy_grid")

print(ModelEvaluationTable)
```

Measurement	Linear SVC Model
Accuracy	0.892384
Error Rate	0.107616
Sensitivity	0.954106
Recall	0.954106
Specificity	0.757895
Precision	0.895692
F1	0.923977
F2	0.941822
F0.5	0.906795

In [68]:
```
data1
```

Out[68]:
```
[['Accuracy', 0.8923841059602649],
 ['Error Rate', 0.10761589403973515],
 ['Sensitivity', 0.9541062801932367],
 ['Recall', 0.9541062801932367],
 ['Specificity', 0.7578947368421053],
 ['Precision', 0.8956916099773242],
 ['F1', 0.9239766081871345],
 ['F2', 0.9418216499761562],
 ['F0.5', 0.9067952249770431]]
```

In [69]:
```
Data_metric_results_TheHill=pd.DataFrame(data1)
Data_metric_results_TheHill.head()
```

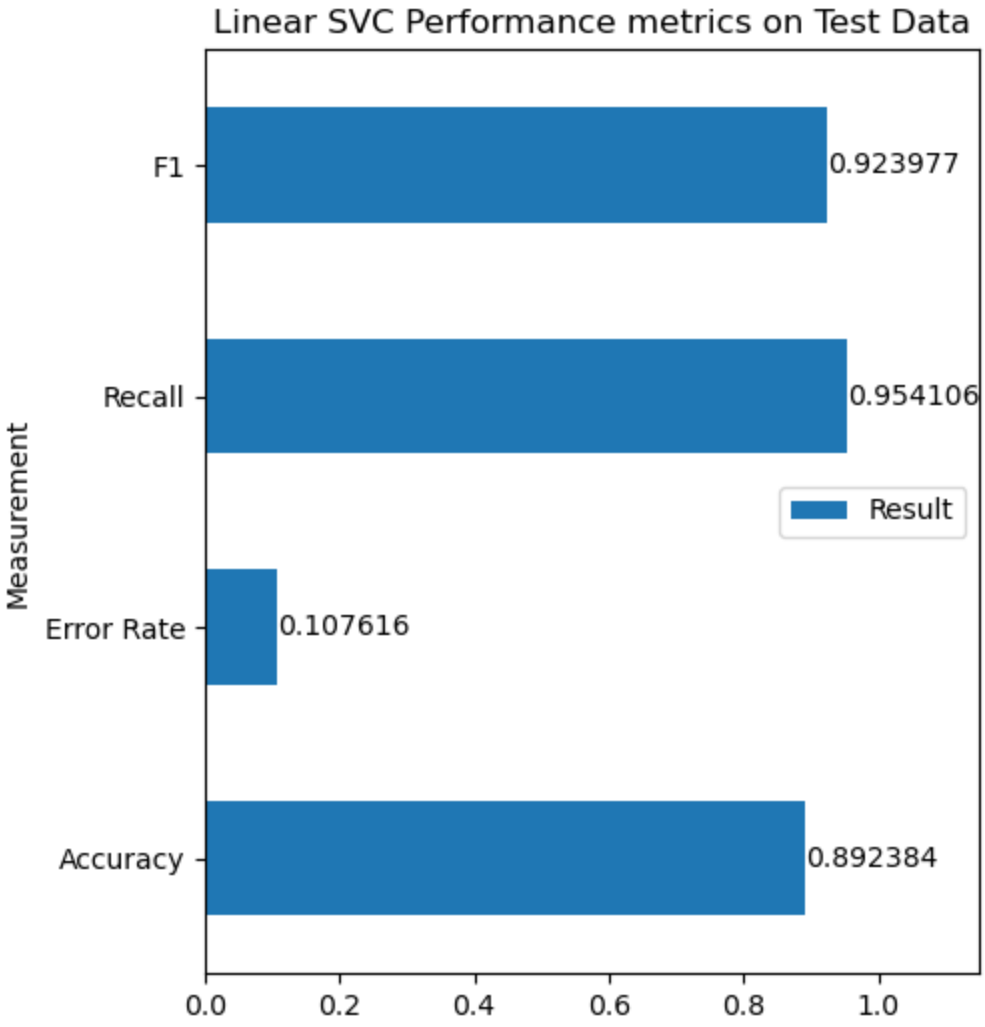
Out[69]:

	0	1
0	Accuracy	0.8924
1	Error Rate	0.1076
2	Sensitivity	0.9541
3	Recall	0.9541
4	Specificity	0.7579

In [70]:
```
Data_metric_results_TheHill.rename (columns = {0:'Measurement'}, inplace=True)
Data_metric_results_TheHill.rename (columns = {1:'Result'}, inplace=True)
```

In [71]:
```
#plt.bar(x=ModelEvaluationTable)
ax=Data_metric_results_TheHill[(Data_metric_results_TheHill['Measurement'] == 'Accu
                               (Data_metric_results_TheHill['Measurement'] == 'Recall'
                               (Data_metric_results_TheHill['Measurement'] == 'F1') |
                               (Data_metric_results_TheHill['Measurement'] == 'Error R

                                     figsize=(5,6),
                                     title='Linear SVC Performance metrics on Test Dat
ax.bar_label(ax.containers[0])
ax.set_xlim(right=1.15)
```

Out[71]:
```
(0.0, 1.15)
```



Business problem application

In [72]:
```
center_df01 = pd.read_csv(file_in_path02)

print(center_df01.shape)
display(center_df01.head())
```

```
(181, 12)
```

	Unnamed: 0.1	Unnamed: 0	Source	Author	Title	
0	0	0	The Hill	Zach Schonfeld	Ketanji Brown Jackson issues solo dissent in r...	https://thehill.com/regulation/cou
1	1	1	The Hill	Brett Samuels	How Biden pulled it off...	https://thehill.com/homenews/admir
2	2	2	The Hill	the hill	How Christie could be wildcard in 2024 race...	https://thehill.com/homenews/campaig
3	3	3	The Hill	Alexander Bolton	Schumer announces agreement to pass debt ceili...	https://thehill.com/homenews/senate/
4	4	4	The Hill	Zack Budryk	Kaine introduces amendment to strip Manchin-ba...	https://thehill.com/policy/energy-er

In [73]:
```
# Apply transformers to pandas dataframe, w/ new col containing tokens
center_df01['processed_text'] = center_df01['article_text']\
.progress_apply(prepare, pipeline=transformers01)

center_df01['processed_text_split'] = center_df01['processed_text']\
.progress_apply(str.split)

center_df01['num_tokens'] = center_df01['processed_text_split']\
.map(len)

display(center_df01.head())

# Review unique tokens across entire dataset
for c in range(0,1):
    try:
        print(center_df01['processed_text'][c], '\n')
    except:
        print(f'Skip {c}')
```

```
100%|██████████| 181/181 [00:01<00:00, 128.38it/s]
100%|██████████| 181/181 [00:00<00:00, 25704.92it/s]
```

	Unnamed: 0.1	Unnamed: 0	Source	Author	Title	
0	0	0	The Hill	Zach Schonfeld	Ketanji Brown Jackson issues solo dissent in r...	https://thehill.com/regulation/cou
1	1	1	The Hill	Brett Samuels	How Biden pulled it off...	https://thehill.com/homenews/admin
2	2	2	The Hill	the hill	How Christie could be wildcard in 2024 race...	https://thehill.com/homenews/campaig
3	3	3	The Hill	Alexander Bolton	Schumer announces agreement to pass debt ceili...	https://thehill.com/homenews/senate/
4	4	4	The Hill	Zack Budryk	Kaine introduces amendment to strip Manchin-ba...	https://thehill.com/policy/energy-en

```
liberal justice ketanji brown jackson issued first solo dissent supreme court merits
case thursday disagreeing colleagues labor dispute ruling makes easier companies sue
worker strikes 8 1 decision high court overturned lower ruling found federal union l
aws preempted concrete company glacier northwest bringing lawsuit international brot
herhood teamsters represents companys truck drivers jackson wrote courts no business
delving particular labor dispute time the majority also misapplies boards cases mann
er threatens impede boards uniform development labor law erode right strike jackson
dissented case arose union directed drivers go strike morning company mixing concret
e loading onto trucks making deliveries concrete mixed day ruined glacier sued union
damages state court 1959 supreme court precedent san diego building trades council v
garmon national labor relations act nlra federal law governs strikes collective barg
aining preempts state law two arguably conflict union got lawsuit tossed washington
supreme court garmon glacier northwest appealed nations highest court majority opini
on authored conservative justice amy coney barrett joined four colleagues court rule
d nlra preempt lawsuit strike take reasonable precautions protect companys property
foreseeable imminent danger the unions actions resulted destruction concrete glacier
prepared day also posed risk foreseeable aggravated imminent harm glaciers trucks un
ion took affirmative steps endanger glaciers property rather reasonable precautions
mitigate risk nlra arguably protect conduct barrett wrote three additional conservat
ive justices justices samuel alito clarence thomas neil gorsuch wrote separately rev
erse unions win grounds jackson hand stood alone dissenting marking first solo disse
nt merits case since joining bench last year jackson has however dissented solo outs
ide courts normal docket jpmorgan says former virgin islands first lady aided epstei
n trump indictment lays bare security risks storage mar lago noted complaint nationa
l labor relations boards general counsel filed state supreme courts ruling alleged g
lacier northwest engaged unfair labor practices relation strike majority found this
issue properly us lower courts addressed significance complaint leaving state courts
consider case proceeds the filing general counsels administrative complaint necessar
ily suffices establish unions strike conduct arguably protected within meaning garmo
n thus general counsels complaint marked end court involvement matter jackson wrote
```

In [74]:
```
nlm_apply_x01_mtx = nlm_tfidf.transform(center_df01['processed_text'])

print(nlm_apply_x01_mtx.shape)
display(nlm_apply_x01_mtx)
```

```
(181, 48932)
<181x48932 sparse matrix of type '<class 'numpy.float64'>'
        with 64886 stored elements in Compressed Sparse Row format>
```

In [75]:
```
display_samp_dwm(sm=nlm_apply_x01_mtx,
                 vec=nlm_tfidf,
                 n=(17,11),
                 rs_tup=(5,1699))
```

	nixon	follow london	efforts block	federal employees	request meeting	sexual assaults	chief executive	almost half	senior fellow	dee santi
27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0000	0.0
152	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0000	0.0
154	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0000	0.0
131	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0000	0.0
75	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0000	0.0
77	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0000	0.0
170	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0222	0.0
121	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0000	0.0
43	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0000	0.0
48	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0000	0.0
53	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0000	0.0

Out[75]:
```
array(['03', '04', '05', ..., 'zoos', 'zucker', 'zuckerberg'],
      dtype=object)
```

In [76]:
```
nlm_apply_mtx_pred_prob = m2v1_gbc.predict_proba(nlm_apply_x01_mtx)

print(nlm_apply_mtx_pred_prob.shape)
print(nlm_apply_mtx_pred_prob[:10])

nlm_apply_mtx_pred = m2v1_gbc.predict(nlm_apply_x01_mtx)

print(nlm_apply_mtx_pred.shape)
print(nlm_apply_mtx_pred)
```

```
(181, 2)
[[0.1943 0.8057]
 [0.361  0.639 ]
 [0.9765 0.0235]
 [0.0586 0.9414]
 [0.8779 0.1221]
 [0.7489 0.2511]
 [0.0426 0.9574]
 [0.8892 0.1108]
 [0.87   0.13  ]
 [0.     1.    ]]
(181,)
['right' 'right' 'left' 'right' 'left' 'left' 'right' 'left' 'left'
 'right' 'left' 'left' 'right' 'right' 'left' 'right' 'left' 'right'
 'left' 'right' 'right' 'left' 'right' 'right' 'right' 'right' 'right'
 'left' 'right' 'right' 'left' 'left' 'right' 'left' 'left' 'right'
 'right' 'right' 'right' 'right' 'left' 'left' 'right' 'left' 'left'
 'right' 'right' 'left' 'right' 'left' 'right' 'right' 'right' 'right'
 'right' 'right' 'left' 'right' 'left' 'right' 'left' 'left' 'right'
 'left' 'right' 'left' 'left' 'right' 'right' 'right' 'right' 'right'
 'right' 'right' 'right' 'right' 'right' 'right' 'right' 'right' 'right'
 'left' 'right' 'right' 'left' 'left' 'left' 'left' 'right' 'right' 'left'
 'right' 'right' 'right' 'right' 'right' 'left' 'left' 'right' 'left'
 'left' 'right' 'left' 'left' 'right' 'left' 'left' 'right' 'left' 'left'
 'right' 'right' 'right' 'right' 'left' 'right' 'right' 'right' 'left'
 'right' 'right' 'right' 'left' 'right' 'left' 'right' 'right' 'left'
 'left' 'right' 'left' 'right' 'right' 'right' 'left' 'right' 'right'
 'right' 'left' 'left' 'left' 'left' 'left' 'left' 'right' 'right' 'right'
 'right' 'right' 'left' 'right' 'left' 'right' 'right' 'right' 'left'
 'right' 'left' 'right' 'right' 'right' 'right' 'left' 'right' 'right'
 'right' 'left' 'left' 'right' 'right' 'right' 'right' 'right' 'right'
 'left' 'right' 'right' 'right' 'right' 'right' 'right']
```

In [77]:
```
# Compute the maximum values along the second dimension
max_values = np.amax(nlm_apply_mtx_pred_prob, axis=1)
max_values_df01 = pd.DataFrame(max_values,
                               columns=['decision_prob'])
max_values_df01['pred'] = nlm_apply_mtx_pred
print(max_values_df01.shape)
display(max_values_df01.head())
```

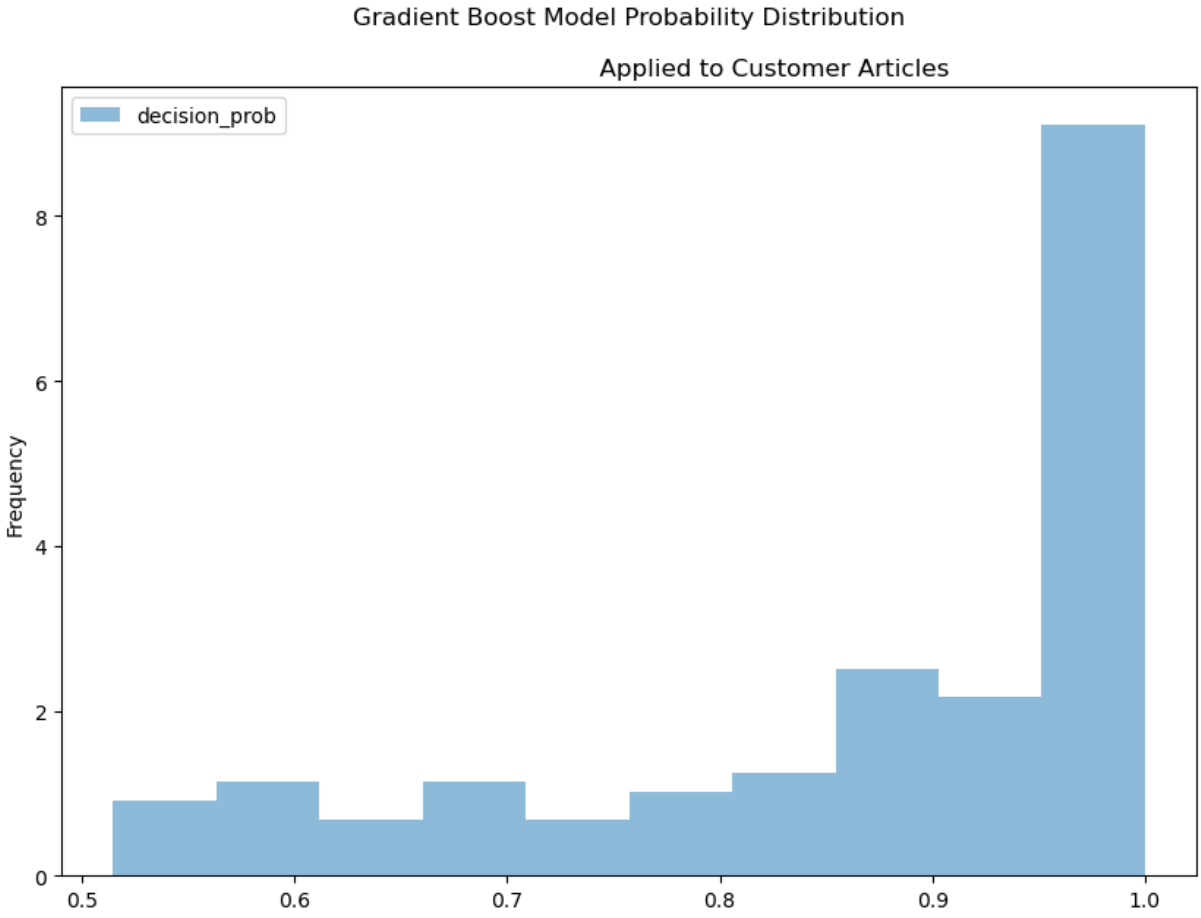
```
(181, 2)
```

	decision_prob	pred
0	0.8057	right
1	0.6390	right
2	0.9765	left
3	0.9414	right
4	0.8779	left

In [78]:
```
max_values_df01['decision_prob'].plot(kind="hist",
                                      density=True,
```

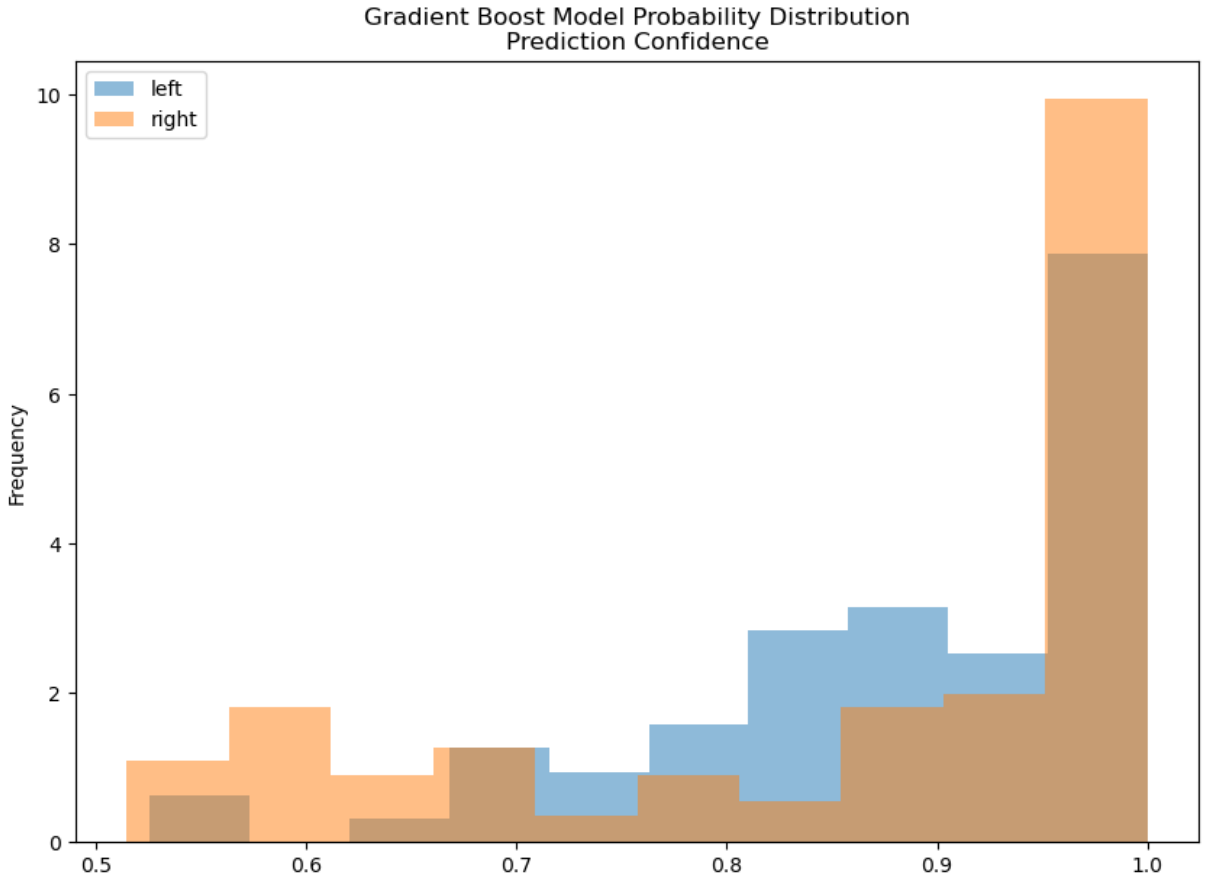
```
                                        alpha=0.5,
                                        legend=True,
                                        figsize=(10,7),
title='''Gradient Boost Model Probability Distribution\n
                                        Applied to Customer Articles''')
```

Out[78]:
```
<Axes: title={'center': 'Gradient Boost Model Probability Distribution\n\n
Applied to Customer Articles'}, ylabel='Frequency'>
```



In [79]:
```
max_values_df01.groupby('pred')['decision_prob'].plot(kind="hist",
                                                      density=True,
                                        alpha=0.5,
                                        legend=True,
                                        figsize=(10,7),
                            title='''Gradient Boost Model Probability Distribution
Prediction Confidence''')
```

Out[79]:
```
pred
left     Axes(0.125,0.11;0.775x0.77)
right    Axes(0.125,0.11;0.775x0.77)
Name: decision_prob, dtype: object
```



In [80]:
```
max_values_df02 = pd.DataFrame(nlm_apply_mtx_pred_prob.round(4),
                               columns=['left', 'right'])
max_values_df02['pred'] = nlm_apply_mtx_pred
max_values_df02
```

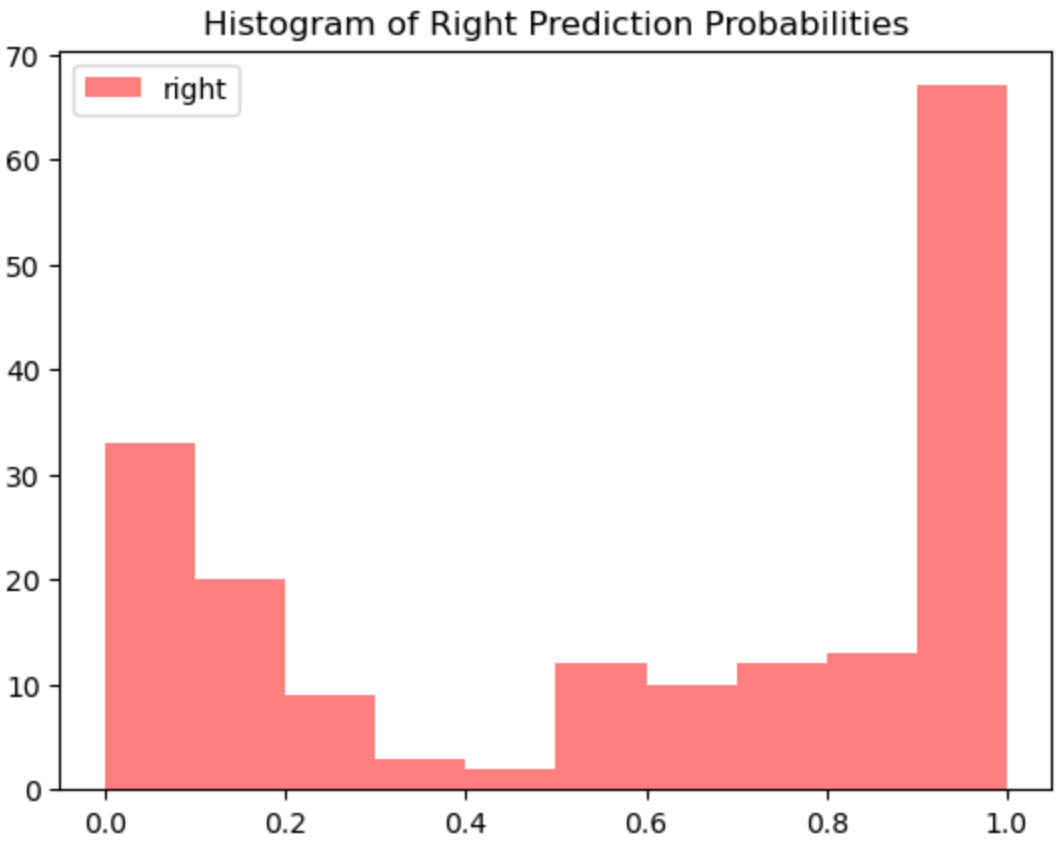
Out[80]:

	left	right	pred
0	0.1943	0.8057	right
1	0.3610	0.6390	right
2	0.9765	0.0235	left
3	0.0586	0.9414	right
4	0.8779	0.1221	left
...	...	...	...
176	0.4365	0.5635	right
177	0.0708	0.9292	right
178	0.0008	0.9992	right
179	0.4365	0.5635	right
180	0.0063	0.9937	right

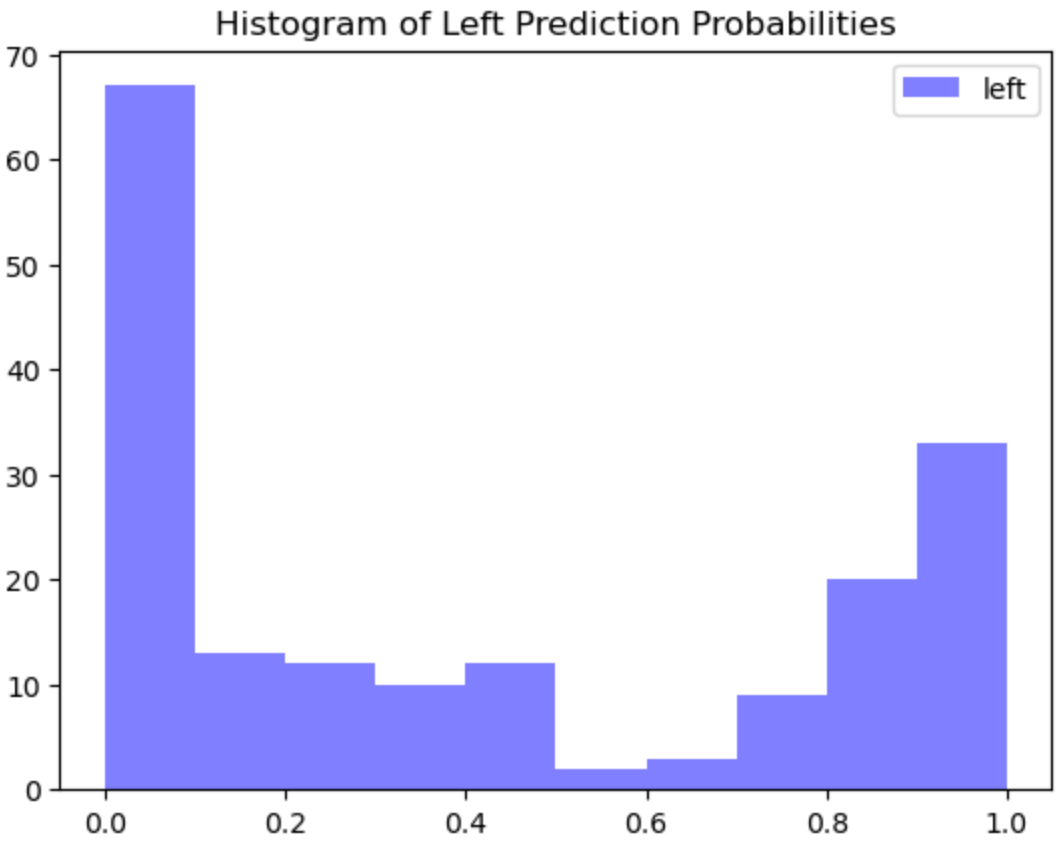
181 rows × 3 columns

In [81]:

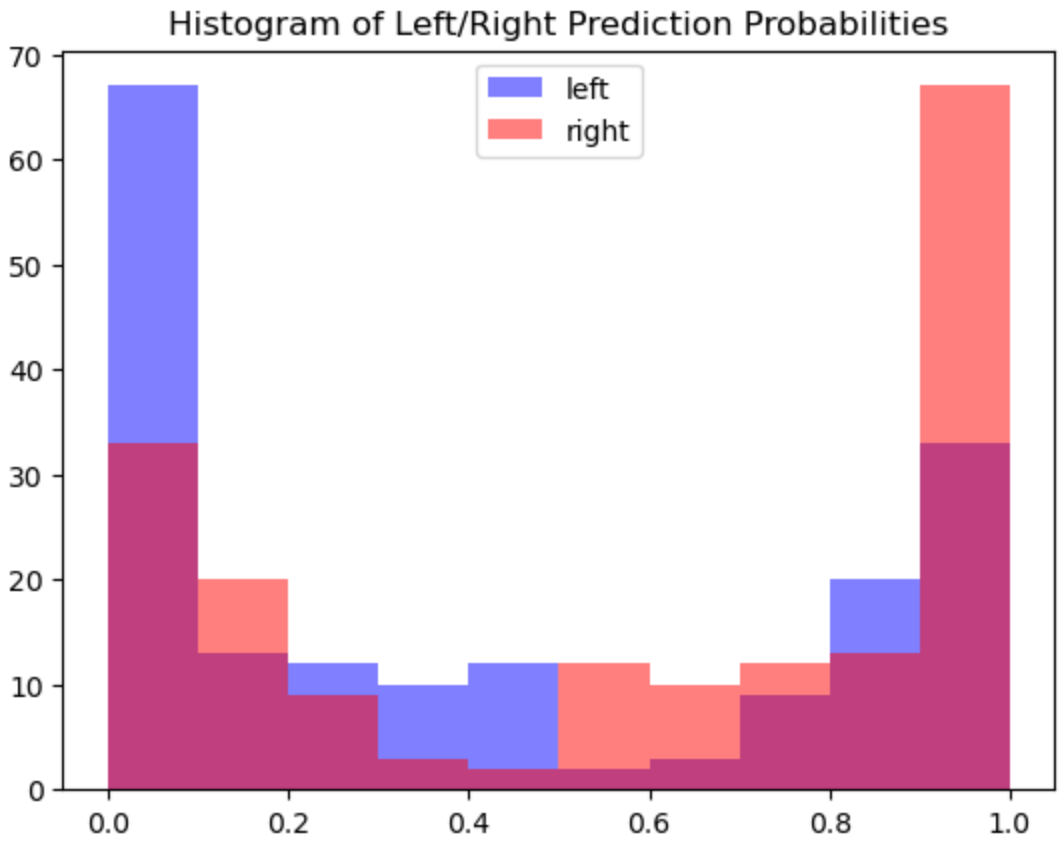
```
# Plotting histograms
#plt.hist(max_values_df02['left'], bins=10, alpha=0.5, color='blue', label='Column
plt.hist(max_values_df02['right'], bins=10, alpha=0.5, color='red',
         label='right')

# Adding legend and title
plt.legend()
plt.title('Histogram of Right Prediction Probabilities')

# Displaying the plot
plt.show()
```



In [82]:
```
# Plotting histograms
#plt.hist(max_values_df02['left'], bins=10, alpha=0.5, color='blue', label='Column
plt.hist(max_values_df02['left'], bins=10, alpha=0.5, color='blue',
         label='left')

# Adding legend and title
plt.legend()
plt.title('Histogram of Left Prediction Probabilities')

# Displaying the plot
plt.show()
```



In [83]:
```
# Plotting histograms
plt.hist(max_values_df02['left'], bins=10,
         alpha=0.5, color='blue', label='left')
plt.hist(max_values_df02['right'], bins=10,
         alpha=0.5, color='red', label='right')

# Adding legend and title
plt.legend()
plt.title('Histogram of Left/Right Prediction Probabilities')

# Displaying the plot
plt.show()
```


Histogram of Left/Right Prediction Probabilities
left
right
70
60
50
40
30
20
10
0
0.0
0.2
0.4
0.6
0.8
1.0