Illuminating Mainstream Media Political Bias through Text Mining

Aaron Carr, Azucena Faus, and Dave Friesen - ADS-509-01-SU23

In [1]:
```
__author__ = 'Aaron Carr, Azucena Faus, Dave Friesen'
__email__ = 'acarr@sandiego.edu, afaus@sandiego.edu, dfriesen@sandiego.edu'
__version__ = '1.0'
__date__ = 'June 2023'
```

Setup

In [2]:
```
# Import basic and data access libraries
import numpy as np
import pandas as pd
from profiler import profile, profile_cat

# Import pre-processing, model and performance evaluation libraries
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from nltk.tokenize import sent_tokenize, word_tokenize
from sklearn.decomposition import LatentDirichletAllocation
from model_process import ModelProcess

# Import lexicons
#import nltk
#nltk.download('opinion_lexicon')
from nltk.corpus import opinion_lexicon

# Import visualization libraries
from matplotlib import pyplot as plt
%matplotlib inline
import seaborn as sns
from wordcloud import WordCloud

# Import utility libraries
from collections import Counter, defaultdict
from tqdm import tqdm; tqdm.pandas()
```

In [3]:
```
# Set basic np, pd, and plt output defaults (keeping this code 'clean')
%run -i 'defaults.py'
```

Data Ingestion

In [4]:
```
# Instantiate and confirm master dataframe
master_df = pd.read_csv('../data/master.csv')
master_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4509 entries, 0 to 4508
Data columns (total 7 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   source_name   4509 non-null   object
```

```
 1   author         4472 non-null   object
 2   title          4509 non-null   object
 3   url            4509 non-null   object
 4   publish_date   4509 non-null   object
 5   content        1158 non-null   object
 6   article_text   4508 non-null   object
dtypes: object(7)
memory usage: 246.7+ KB
```

In [5]:

```
# "Blanket" label data based on purported source leaning
target_cls_col = 'lean'

def assign_lean(source_name):
    if source_name == 'Breitbart News' or source_name == 'Fox News':
        return 'right'
    elif source_name == 'CNN' or source_name == 'The Washington Post':
        return 'left'
    else:
        return np.nan
master_df[target_cls_col] = master_df['source_name'].apply(lambda x: assign_lean(x))
```

Tokenization and Cleaning

In [6]:

```
# Stopword removal function, with related initialization
from nltk.corpus import stopwords
sw = stopwords.words('english')
def remove_stop(tokens):
    filtered_tokens = [word for word in tokens if word not in sw]
    return(filtered_tokens)

# Token join back to string
def join_tokens(tokens):
    return ' '.join(tokens)

# Tokenizing function
def tokenize(text):
    return(text.split())  # Tokenize on white space

# Emoji-to-text conversion function
import emoji
def convert_emojis(text):
#    return emoji.demojize(text)
    return emoji.demojize(text).replace('_', ' ')

# Contains-emojis function, with related initialization
all_language_emojis = set()
for country in emoji.EMOJI_DATA :
    for em in emoji.EMOJI_DATA[country]:
        all_language_emojis.add(em)
def contains_emoji(s):
    s = str(s)
    emojis = [ch for ch in s if ch in all_language_emojis]
    return(len(emojis) > 0)

# Punctuation removal function, with related initialization
from string import punctuation
tw_punct = set(punctuation + '’') - {'#'}
def remove_punct(text, punct_set=tw_punct):
    return(''.join([ch for ch in text if ch not in punct_set]))

# Preparation (pipeline) function
def prepare(text, pipeline):
    tokens = str(text)
```

```
    for transform in pipeline:
        tokens = transform(tokens)
    return(tokens)
```

In [7]:

```
# Set pipeline
pipeline = [str.lower, remove_punct, convert_emojis, tokenize, remove_stop]

# Clean and tokenize master dataframe
master_df['article_tokens'] = master_df['article_text'].progress_apply(lambda x: prepare(x
master_df['article_text_tokenized'] = master_df['article_tokens'].progress_apply(lambda x:
print(master_df['article_tokens'])
print(master_df['article_text_tokenized'])
```

```
100%|██████████████████████████████████| 4509/4509 [00:09<00:00, 493.88it/s]
100%|██████████████████████████████████| 4509/4509 [00:00<00:00, 86571.74it/s]
0       [travelers, alabama, driving, interstate, 65, ...
1       [federal, prosecutor, may, nearing, decision, ...
2       [federal, appeals, court, tuesday, cleared, wa...
3       [speaking, orlando, november, 2015, republican...
4                                                   [nan]
                              ...
4504    [germanys, populist, alternative, germany, afd...
4505    [president, bidens, justice, department, seemi...
4506    [incumbent, turkish, president, recep, tayyip,...
4507    [throughout, month, may, farleft, cnn, attract...
4508    [disney, known, fighting, antigrooming, legisl...
Name: article_tokens, Length: 4509, dtype: object
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                                                     nan
                              ...
4504    germanys populist alternative germany afd surg...
4505    president bidens justice department seemingly ...
4506    incumbent turkish president recep tayyip erdog...
4507    throughout month may farleft cnn attracted mea...
4508    disney known fighting antigrooming legislation...
Name: article_text_tokenized, Length: 4509, dtype: object
```

Descriptive Stats

In [8]:

```
# Descriptive stats function
def descriptive_stats(tokens, num_tokens=5, verbose=False):
    num_tokens = len(tokens)
    num_unique_tokens = len(set(tokens))  # set() creates unordered set of unique elements
    num_characters = sum(len(token) for token in tokens)  # Finds characters sans spaces
    lexical_diversity = num_unique_tokens / num_tokens

    if verbose:
        print(f'There are {num_tokens} tokens in the data.')
        print(f'There are {num_unique_tokens} unique tokens in the data.')
        print(f'There are {num_characters} characters in the data.')
        print(f'The lexical diversity is {lexical_diversity:.3f} in the data.')

    return([num_tokens, num_unique_tokens, lexical_diversity, num_characters])
```

In [9]:

```
# Descriptive stats across all sources
descriptive_stats([token for sublist in master_df['article_tokens'] for token in sublist])
```

Out[9]: [1977106, 84569, 0.0427741355294051, 12724251]

In [10]:
```
# Standard dataframe profile for confirmation
profile(master_df)
```

100%|██████████| 4509/4509 [00:00<00:00, 1305454.32it/s]

	Dtype	count	unique	na	na%	mean	std	min	max	skew(>=3)	<v0.01	V
source_name	object	4509.0	4.0									
author	object	4472.0	956.0	37.0	0.8							
title	object	4509.0	4509.0									
url	object	4509.0	4509.0									
publish_date	object	4509.0	4487.0									
content	object	1158.0	1158.0	3351.0	74.3							
article_text	object	4508.0	4508.0	1.0								
lean	object	4509.0	2.0									
article_tokens	object	1977106.0	84569.0									
article_text_tokenized	object	4509.0	4509.0									

In [11]:
```
# Descriptive stats aggregating function
def aggregate_and_describe(group):
    aggregate_tokens = [token for sublist in group['article_tokens'].tolist() for token in
    return descriptive_stats(aggregate_tokens)

# Aggregate descriptive stats by source; convert to dataframe; sort and output
grouped_stats = master_df.groupby('source_name').apply(aggregate_and_describe)
grouped_stats_df = pd.DataFrame(grouped_stats.tolist(), index=grouped_stats.index,
                                columns=['num_tokens', 'num_unique_tokens', 'lexical_diver
grouped_stats_df = grouped_stats_df.sort_index(ascending=False)
print(grouped_stats_df)
```

source_name	num_tokens	num_unique_tokens	lexical_diversity	num_characters
The Washington Post	366707	32341	0.09	2370171
Fox News	828739	47097	0.06	5326935
CNN	409422	34724	0.08	2628951
Breitbart News	372238	36815	0.10	2398194

Word Cloud

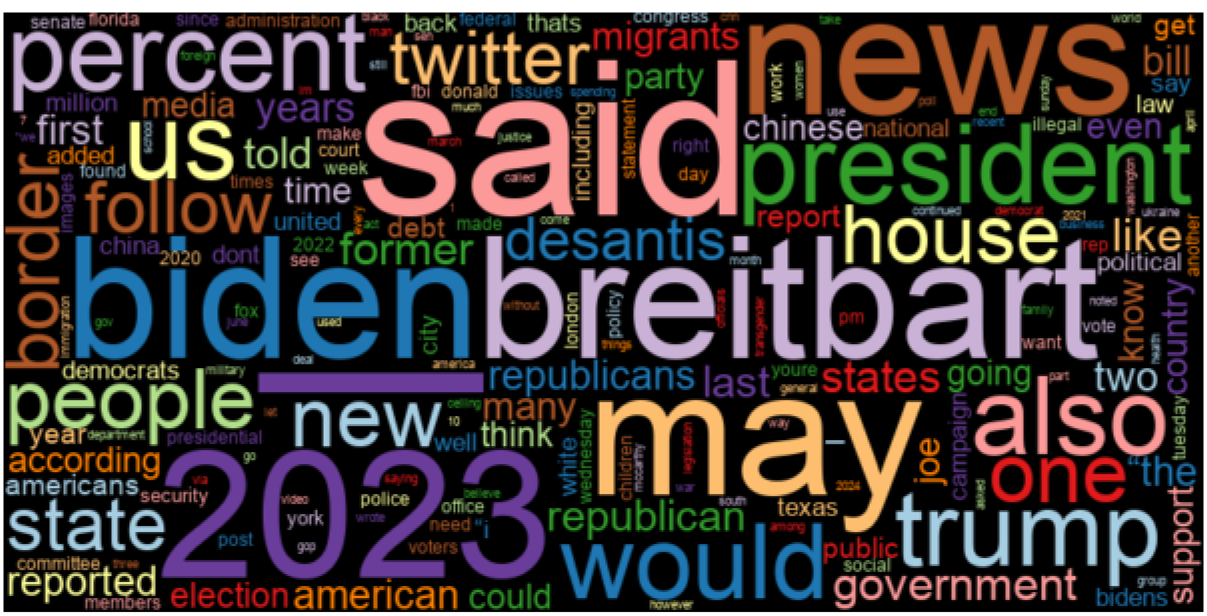
In [12]:
```
# Word cloud function
def wordcloud(word_freq, title=None, max_words=200, stopwords=None):
    wc = WordCloud(font_path='/Library/Fonts/Arial.ttf',
                   width=800, height=400,
                   background_color= "black", colormap="Paired",
                   max_font_size=150, max_words=max_words)

    # Convert data frame into dict
    if type(word_freq) == pd.Series:
        counter = Counter(word_freq.fillna(0).to_dict())
    else:
```

```
        counter = word_freq

    # filter stop words in frequency counter
    if stopwords is not None:
        counter = {token:freq for (token, freq) in counter.items()
                              if token not in stopwords}
    wc.generate_from_frequencies(counter)

    plt.title(title)

    plt.imshow(wc, interpolation='bilinear')
    plt.axis("off")

    plt.show()

# Word count function counter
def count_words(df, column='article_tokens', preprocess=None, min_freq=2):
    # Process tokens and update counter
    def update(doc):
        tokens = doc if preprocess is None else preprocess(doc)
        counter.update(tokens)

    # Create counter and run through all data
    counter = Counter()
    df[column].map(update)

    # Transform counter into data frame
    freq_df = pd.DataFrame.from_dict(counter, orient='index', columns=['freq'])
    freq_df = freq_df.query('freq >= @min_freq')
    freq_df.index.name = 'token'

    return freq_df.sort_values('freq', ascending=False)
```

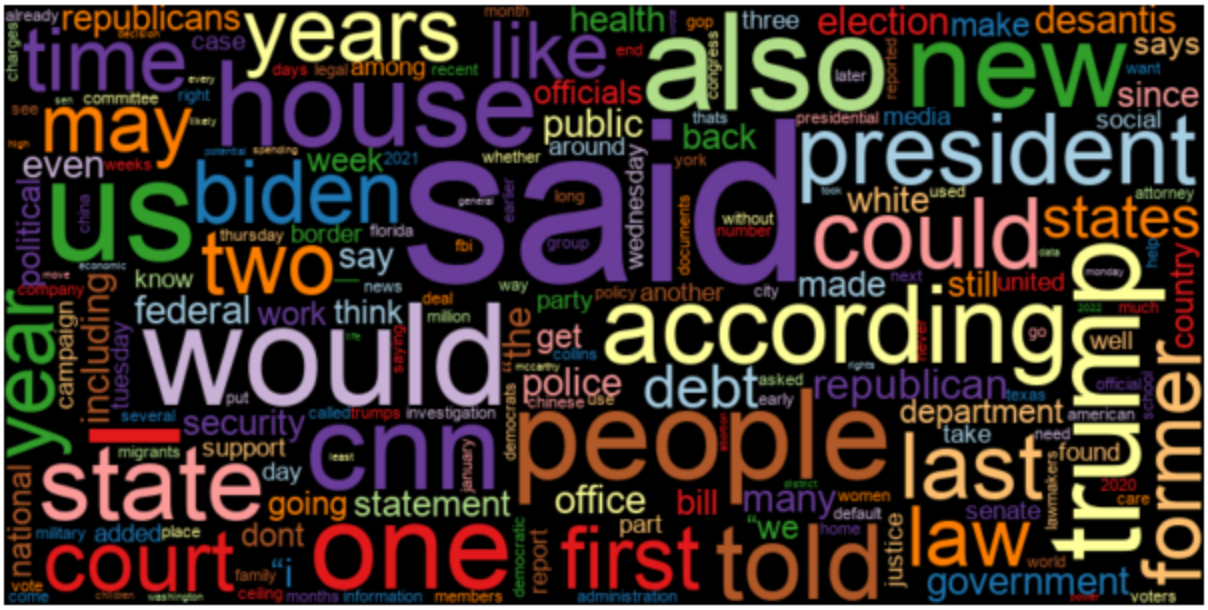
In [13]:

```
# Iterate and produce word cloud by source
for name, group in master_df.groupby('source_name'):
    print(f"Wordcloud for source: {name}")
    wordcloud(count_words(group)['freq'].to_dict())
```

```
Wordcloud for source: Breitbart News
```

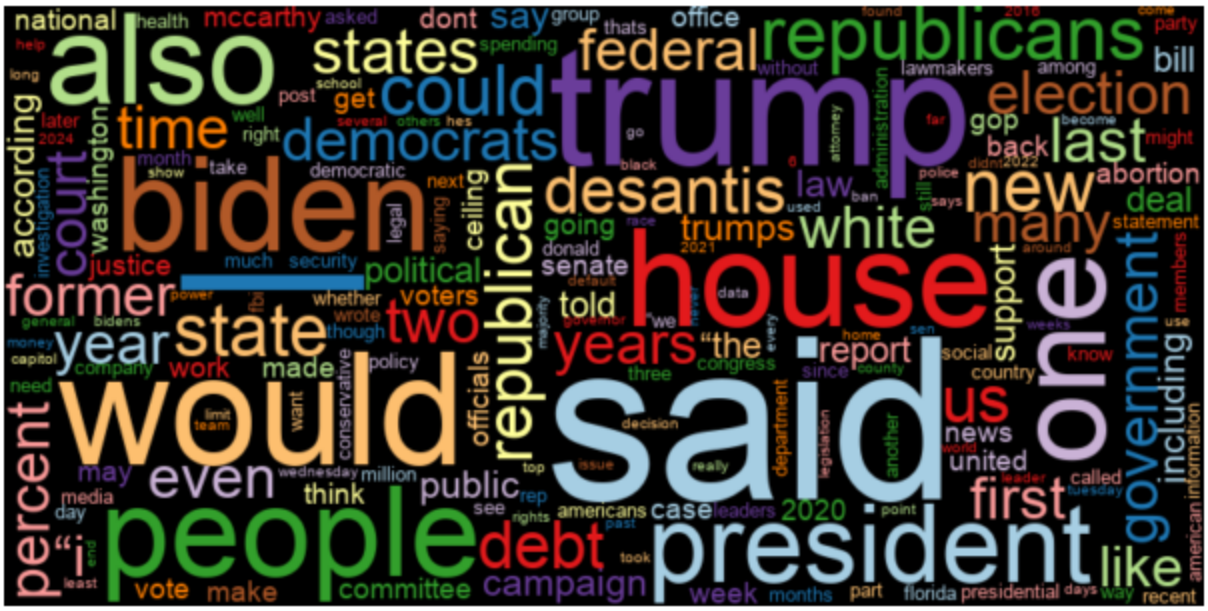



```
Wordcloud for source: CNN
```



Wordcloud for source: Fox News

Wordcloud for source: The Washington Post



Partition Data

In [14]:
```
# Set splits
train_ratio = 0.7; val_ratio = 0.20; test_ratio = 0.10

# Split and profile
train_df, test_df = train_test_split(master_df, test_size=1-train_ratio,
                                     random_state=42, stratify=master_df[target_cls_col])
val_df, test_df = train_test_split(test_df, test_size=test_ratio/(test_ratio+val_ratio),
                                   random_state=42, stratify=test_df[target_cls_col])
profile_cat(train_df, [target_cls_col])
```

```
lean -
right 71.17
 left 28.83
```

Topic Modeling

In [15]:
```
# Topic summarization function, from BTAP repo
def display_topics(model, features, no_top_words=5):
    for topic, words in enumerate(model.components_):
        total = words.sum()
        largest = words.argsort()[::-1]  # invert sort order

        print('\nTopic %02d' % topic, end=':')

        out = []
        for i in range(0, no_top_words):
            out.append(' %s (%2.2f)' % (features[largest[i]], abs(words[largest[i]]*100.0/
        print(';'.join(out), end='')
```

In [16]:
```
# Model topics by source
for source_name, group in train_df.groupby('source_name'):
    print(f'Topic modeling for source: {source_name}')

    # Transform article tokens into bag-of-words document-term sparse matrix
    count_vectorizer = CountVectorizer(min_df=0.05, max_df=0.75)
    count_vectors = count_vectorizer.fit_transform(group['article_text_tokenized'])
#    print('Vector shape:', count_vectors.shape)

    lda_model = LatentDirichletAllocation(n_components=5, random_state=42)
    W_lda_matrix = lda_model.fit_transform(count_vectors)
    H_lda_matrix = lda_model.components_

    display_topics(lda_model, count_vectorizer.get_feature_names_out())
    print('\n')
```

```
Topic modeling for source: Breitbart News

Topic 00: percent (5.31); desantis (2.53); trump (2.50); news (1.30); president (1.29)
Topic 01: biden (2.85); border (1.88); house (1.82); migrants (1.29); debt (1.23)
Topic 02: chinese (0.98); people (0.97); government (0.94); china (0.94); may (0.89)
Topic 03: 2023 (1.45); women (1.06); children (1.05); may (1.02); news (0.95)
Topic 04: trump (2.19); president (1.46); think (1.27); thats (1.14); people (1.03)

Topic modeling for source: CNN

Topic 00: people (1.18); health (0.99); new (0.98); one (0.77); like (0.73)
Topic 01: us (1.97); government (0.81); china (0.71); chinese (0.65); security (0.64)
Topic 02: police (1.67); according (1.33); cnn (1.17); told (1.02); people (0.79)
Topic 03: trump (2.55); desantis (1.70); former (1.29); president (1.16); court (1.08)
Topic 04: house (2.23); debt (1.78); would (1.68); biden (1.12); bill (1.12)
```

```
Topic modeling for source: Fox News

Topic 00: ai (1.14); people (1.09); also (0.85); us (0.80); like (0.77)
Topic 01: trump (2.19); president (1.75); desantis (1.59); former (1.31); campaign (1.06)
Topic 02: biden (2.87); house (2.38); president (1.46); debt (1.14); fbi (1.12)
Topic 03: border (2.02); state (1.98); school (1.41); law (1.38); migrants (1.21)
Topic 04: police (1.77); according (0.92); told (0.90); two (0.79); one (0.78)

Topic modeling for source: The Washington Post

Topic 00: state (1.91); abortion (1.39); republicans (0.94); bill (0.84); ban (0.81)
Topic 01: trump (1.31); court (0.92); election (0.76); case (0.74); justice (0.72)
Topic 02: trump (3.52); desantis (1.91); president (1.06); trumps (0.85); election (0.83)
Topic 03: people (1.09); states (0.68); new (0.63); us (0.56); health (0.54)
Topic 04: house (2.33); debt (2.05); biden (1.77); republicans (1.29); mccarthy (1.19)
```

Text Summarization and Sentiment Analysis

In [17]:

```
# NLTK opinion lexicon
positive_words = set(opinion_lexicon.positive())
negative_words = set(opinion_lexicon.negative())
```

In [18]:

```
# List of "assumed" political phrases
political_phrases = ['gun rights', 'voting rights', 'climate change', 'immigration reform'
                     'tax cuts', 'universal healthcare']
```

In [19]:

```
# Group train_df by 'source_name' for source-level comparison
grouped_df = train_df.groupby('source_name')

# Create dictionaries for scores
political_phrase_scores = {}
sentiment_scores = {}

# Iterate over sources and calc TF-IDF scores vs. political phrases
for source, group in tqdm(grouped_df):
    tfidf_vectorizer = TfidfVectorizer(ngram_range=(1, 3))
    tfidf_vectors = tfidf_vectorizer.fit_transform(group['article_text_tokenized'])

    # Calc TF-IDF sum (scores) where political phrases found
    scores = {}
    sentiment = defaultdict(lambda: defaultdict(int))

    # Iterate over political phrases
    for phrase in political_phrases:
        try:
            index = tfidf_vectorizer.get_feature_names_out().tolist().index(phrase)  # try
            scores[phrase] = tfidf_vectors[:, index].sum()  # and sum related score
        except ValueError:
            pass  # didn't find political phrase

        # Iterate over each article in the group to calc sentiment
        for text in group['article_text_tokenized']:
            # Tokenize text into sentences because we're calc'ing sentiment on phrase-rele
            sentences = sent_tokenize(text)

            # Check each sentence if it contains the political phrase
            for sentence in sentences:
                if phrase in sentence:
                    # [Tokenize the sentence into words
                    tokens = word_tokenize(sentence)
```

```
                    # Count positive and negative words
                    for word in tokens:
                        if word in positive_words:
                            sentiment[phrase]['positive'] += 1
                        elif word in negative_words:
                            sentiment[phrase]['negative'] += 1

    # Add the scores to the dictionary
    political_phrase_scores[source] = scores
    sentiment_scores[source] = dict(sentiment)
```

```
100%|██████████| 4/4 [00:23<00:00,  5.84s/it]
```

In [20]:

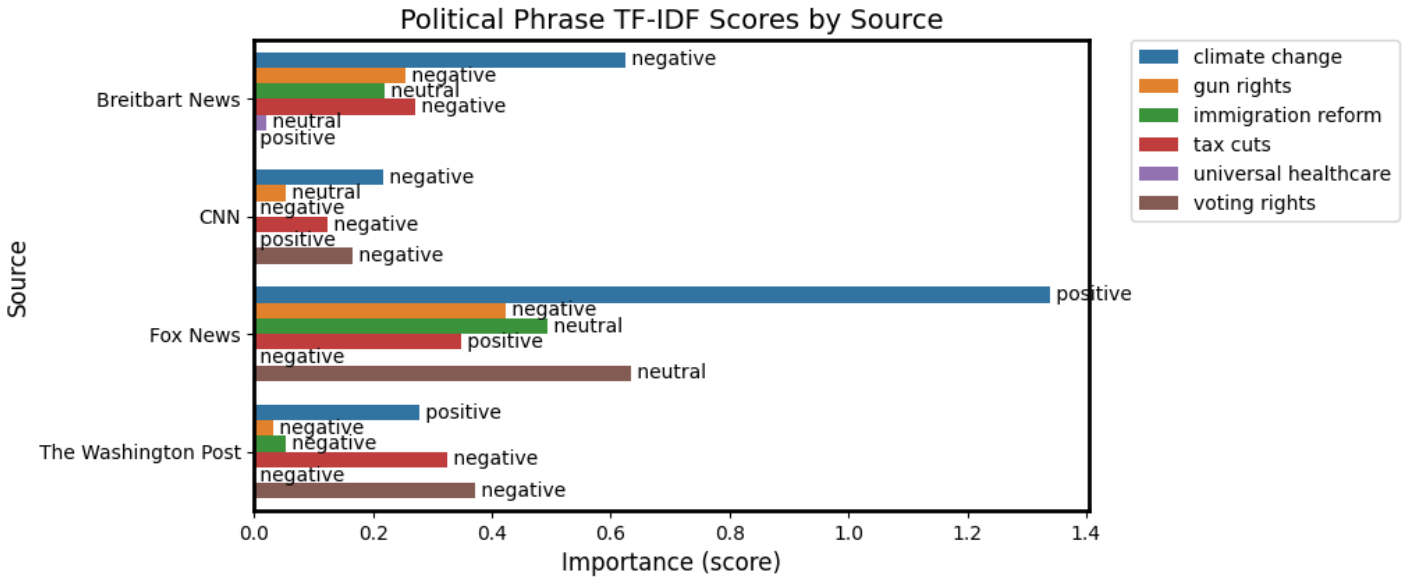
```
# Calc aggregate scores against which to compare "hits" above
all_scores = np.asarray(tfidf_vectors.sum(axis=0)).flatten()
mean_score = np.mean(all_scores)
median_score = np.median(all_scores)

results_df = pd.DataFrame()

# Iterate over sources and political phrase TF-IDF scores and show results
for source_name in political_phrase_scores:
    print(f'\nScores for {source_name}:')
    phrase_scores = political_phrase_scores[source_name]
    sentiment = sentiment_scores[source_name]

    # . . . by political phrase
    results = []
    for phrase in political_phrases:
        score = phrase_scores.get(phrase, 0)
        relative_to_mean = score / mean_score if mean_score != 0 else 0
        relative_to_median = score / median_score if median_score != 0 else 0

        # Categorize based on relative_to_median (otherwise arbitrary)
        if relative_to_median > 10:
            category = 'high'
        elif 5 < relative_to_median <= 10:
            category = 'medium'
        else:
            category = 'low'

        sentiment_phrase = sentiment.get(phrase, {'positive': 0, 'negative': 0})
        results.append({
            'source_name': source_name,
            'phrase': phrase,
            'score': score,
            'relative_to_mean': relative_to_mean,
            'relative_to_median': relative_to_median,
            'category': category,
            'p_sentiment': sentiment_phrase['positive'],
            'n_sentiment': sentiment_phrase['negative'],
            'sentiment': sentiment_phrase['positive'] + (sentiment_phrase['negative'] * -1
        })

    for result in results:
        result['sentiment_label'] = 'positive' if result['sentiment'] > 0 else 'negative'

    # Sort results by score
    results.sort(key=lambda x: x['score'], reverse=True)

    # Print sorted results
    for result in results:
        print(f"{result['phrase']}:  {result['category']} importance ",
              f"{result['sentiment_label']} sentiment")
```

```
    results_df = pd.concat([results_df, pd.DataFrame(results)])
```

```
Scores for Breitbart News:
climate change:  high importance  negative sentiment
tax cuts:  high importance  positive sentiment
gun rights:  high importance  negative sentiment
immigration reform:  high importance  positive sentiment
universal healthcare:  low importance  negative sentiment
voting rights:  low importance  neutral sentiment

Scores for CNN:
climate change:  high importance  negative sentiment
voting rights:  medium importance  negative sentiment
tax cuts:  medium importance  negative sentiment
gun rights:  low importance  negative sentiment
immigration reform:  low importance  neutral sentiment
universal healthcare:  low importance  neutral sentiment

Scores for Fox News:
climate change:  high importance  negative sentiment
voting rights:  high importance  positive sentiment
immigration reform:  high importance  positive sentiment
gun rights:  high importance  negative sentiment
tax cuts:  high importance  negative sentiment
universal healthcare:  low importance  neutral sentiment

Scores for The Washington Post:
voting rights:  high importance  negative sentiment
tax cuts:  high importance  negative sentiment
climate change:  high importance  negative sentiment
immigration reform:  low importance  positive sentiment
gun rights:  low importance  negative sentiment
universal healthcare:  low importance  neutral sentiment
```

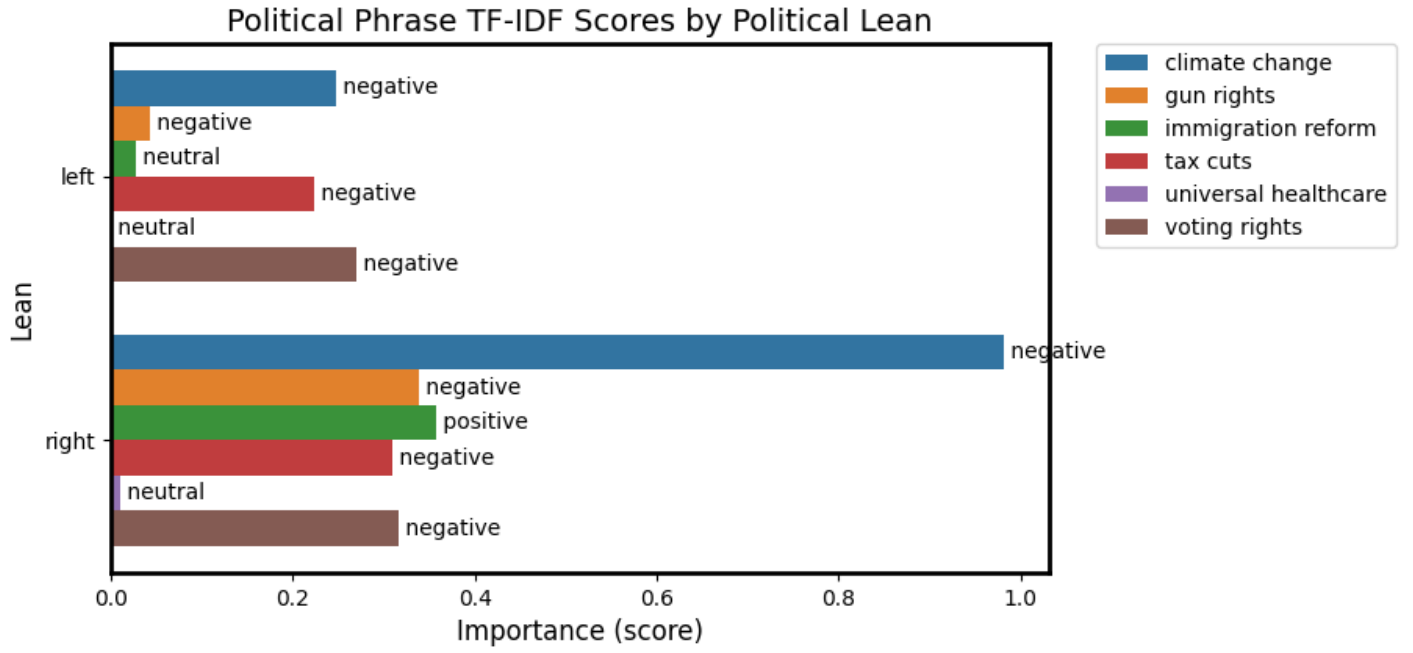
In [21]:

```
# Sort DataFrame by 'source' and 'phrase' to match order of bars in plot
sorted_df = results_df.sort_values(['source_name', 'phrase'])

# Create barplot
fig, ax = plt.subplots()
sns.barplot(data=sorted_df, y='source_name', x='score',
            ax=ax, hue='phrase', errorbar=None)

# Iterate over bars and dataframe rows to add sentiment
for p, (_, row) in zip(ax.patches, sorted_df.iterrows()):
    plt.text(p.get_width(), p.get_y() + p.get_height()/2,
             f' {row["sentiment_label"]}',
             ha='left', va='center')

ax.set_title('Political Phrase TF-IDF Scores by Source')
ax.set_xlabel('Importance (score)')
ax.set_ylabel('Source')

plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)

plt.show()
```



In [22]:

```
# Sort DataFrame by political 'lean' and 'phrase' to match order of bars in plot
lean_df = train_df[['source_name', 'lean']].drop_duplicates()
results_df = pd.merge(results_df, lean_df, on='source_name', how='left')
sorted_df = results_df.sort_values(['lean', 'phrase'])

# Create barplot
fig, ax = plt.subplots()
sns.barplot(data=sorted_df, y='lean', x='score',
            ax=ax, hue='phrase', errorbar=None)

# Iterate over bars and dataframe rows to add sentiment
for p, (_, row) in zip(ax.patches, sorted_df.iterrows()):
    plt.text(p.get_width(), p.get_y() + p.get_height()/2,
             f' {row["sentiment_label"]}',
             ha='left', va='center')

ax.set_title('Political Phrase TF-IDF Scores by Political Lean')
ax.set_xlabel('Importance (score)')
ax.set_ylabel('Lean')

plt.legend(bbox_to_anchor=(1.05, 1), loc=2, borderaxespad=0.)

plt.show()
```



In [ ]: