In [72]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import collections
import langdetect
import spacy
import nltk
import re
import textblob
```

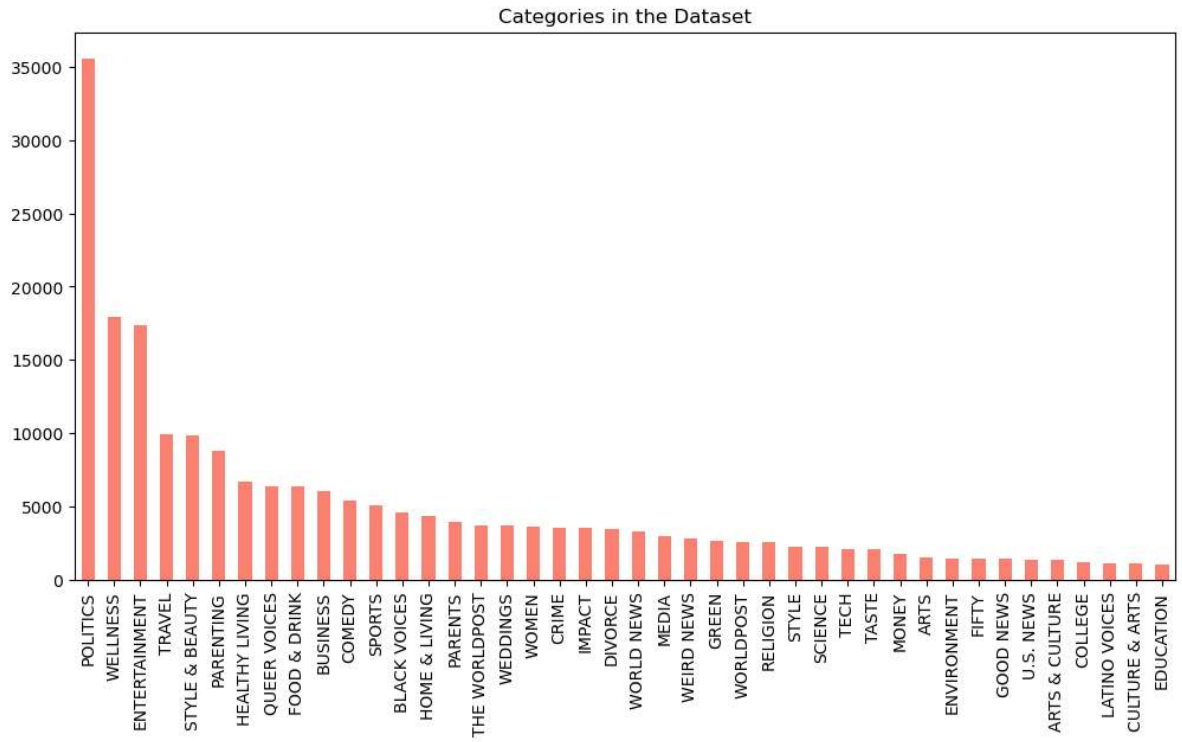
In [2]:
```
df = pd.read_json(r'C:\Users\91900\Downloads\archive\News_Category_Dataset_v3.
```

In [3]:
```
df.head()
```

Out[3]:

	headline	short_description	category
0	Over 4 Million Americans Roll Up Sleeves For O...	Health experts said it is too early to predict...	U.S. NEWS
1	American Airlines Flyer Charged, Banned For Li...	He was subdued by passengers and crew when he ...	U.S. NEWS
2	23 Of The Funniest Tweets About Cats And Dogs ...	"Until you have a dog you don't understand wha...	COMEDY
3	The Funniest Tweets From Parents This Week (Se...	"Accidentally put grown-up toothpaste on my to...	PARENTING
4	Woman Who Called Cops On Black Bird-Watcher Lo...	Amy Cooper accused investment firm Franklin Te...	U.S. NEWS

In [4]:
```
plt.figure(figsize = (12,6))
df['category'].value_counts().plot(kind = 'bar',
                                   color = 'salmon')
plt.title('Categories in the Dataset');
```



In [5]:
```
dtf = df[df["category"].isin(['ENTERTAINMENT','POLITICS','TECH'])]
```

In [6]:
```
dtf.head()
```

Out[6]:

	headline	short_description	category
13	Twitch Bans Gambling Sites After Streamer Scam...	One man's claims that he scammed people on the...	TECH
20	Golden Globes Returning To NBC In January Afte...	For the past 18 months, Hollywood has effectiv...	ENTERTAINMENT
21	Biden Says U.S. Forces Would Defend Taiwan If ...	President issues vow as tensions with China rise.	POLITICS
24	‘Beautiful And Sad At The Same Time’: Ukrainia...	An annual celebration took on a different feel...	POLITICS
28	James Cameron Says He 'Clashed' With Studio Be...	The "Avatar" director said aspects of his 2009...	ENTERTAINMENT

In [7]:
```
dtf.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 55068 entries, 13 to 209522
Data columns (total 3 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   headline           55068 non-null  object
 1   short_description  55068 non-null  object
 2   category           55068 non-null  object
dtypes: object(3)
memory usage: 1.7+ MB
```

In [8]:
```
dtf = dtf.drop('short_description', axis = 1)
dtf.head()
```

Out[8]:

	headline	category
13	Twitch Bans Gambling Sites After Streamer Scam...	TECH
20	Golden Globes Returning To NBC In January Afte...	ENTERTAINMENT
21	Biden Says U.S. Forces Would Defend Taiwan If ...	POLITICS
24	'Beautiful And Sad At The Same Time': Ukrainia...	POLITICS
28	James Cameron Says He 'Clashed' With Studio Be...	ENTERTAINMENT

In [9]:
```
dtf = dtf.rename(columns = {'category':'y', 'headline':'text'})
```

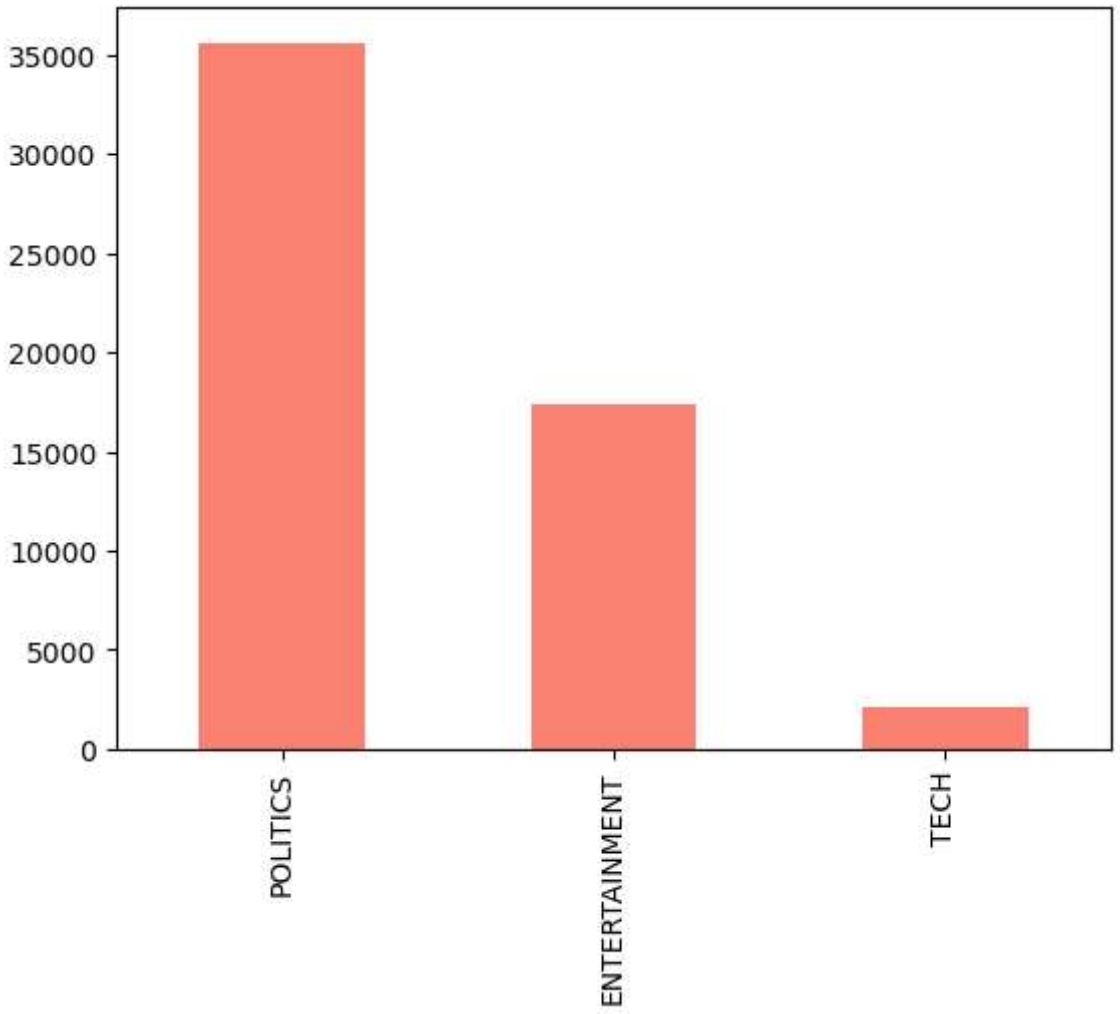
In [12]:
```
dtf.head()
```

Out[12]:

	text	y
13	Twitch Bans Gambling Sites After Streamer Scam...	TECH
20	Golden Globes Returning To NBC In January Afte...	ENTERTAINMENT
21	Biden Says U.S. Forces Would Defend Taiwan If ...	POLITICS
24	'Beautiful And Sad At The Same Time': Ukrainia...	POLITICS
28	James Cameron Says He 'Clashed' With Studio Be...	ENTERTAINMENT

In [11]:
```
dtf['y'].value_counts().plot(kind = 'bar',
                             color = 'salmon')
```

Out[11]: <AxesSubplot:>



In [17]:
```
stopwords = nltk.corpus.stopwords.words("english")
stopwords
```

```
 'ma',
 'mightn',
 "mightn't",
 'mustn',
 "mustn't",
 'needn',
 "needn't",
 'shan',
 "shan't",
 'shouldn',
 "shouldn't",
 'wasn',
 "wasn't",
 'weren',
 "weren't",
 'won',
 "won't",
 'wouldn',
 "wouldn't"]
```

In [27]:
```
def preprocessing(text, stemm = False, lemm = True, stopwords = None):

    #tokenization and removal of stopwords
    text = re.sub(r'[^\w\s]', '', str(text).lower().strip())
    lst_text = text.split()
    if stopwords is not None:
         lst_text = [word for word in lst_text if word not in stopwords]

    #stemming
    if stemm == True:
         ps = nltk.stem.porter.PorterStemmer()
         lst_text = [ps.stem(word) for word in lst_text]

    #lemming
    if lemm == True:
        lem = nltk.stem.wordnet.WordNetLemmatizer()
        lst_text = [lem.lemmatize(word) for word in lst_text]

    text = " ".join(lst_text)
    return text
```

In [43]:
```
dtf["text_clean"] = dtf["text"].apply(lambda x: preprocessing(x, stemm = False
```

In [44]:
```
dtf.head()
```

Out[44]:

	text	y	text_clean
13	Twitch Bans Gambling Sites After Streamer Scam...	TECH	twitch ban gambling site streamer scam folk 20...
20	Golden Globes Returning To NBC In January Afte...	ENTERTAINMENT	golden globe returning nbc january year offair
21	Biden Says U.S. Forces Would Defend Taiwan If ...	POLITICS	biden say u force would defend taiwan china in...
24	‘Beautiful And Sad At The Same Time’: Ukrainia...	POLITICS	beautiful sad time ukrainian cultural festival...
28	James Cameron Says He 'Clashed' With Studio Be...	ENTERTAINMENT	james cameron say clashed studio avatar release

In [45]:
```
print(dtf['text'][13])
dtf['text_clean'][13]
```

```
Twitch Bans Gambling Sites After Streamer Scams Folks Out Of $200,000
```

Out[45]:
```
'twitch ban gambling site streamer scam folk 200000'
```

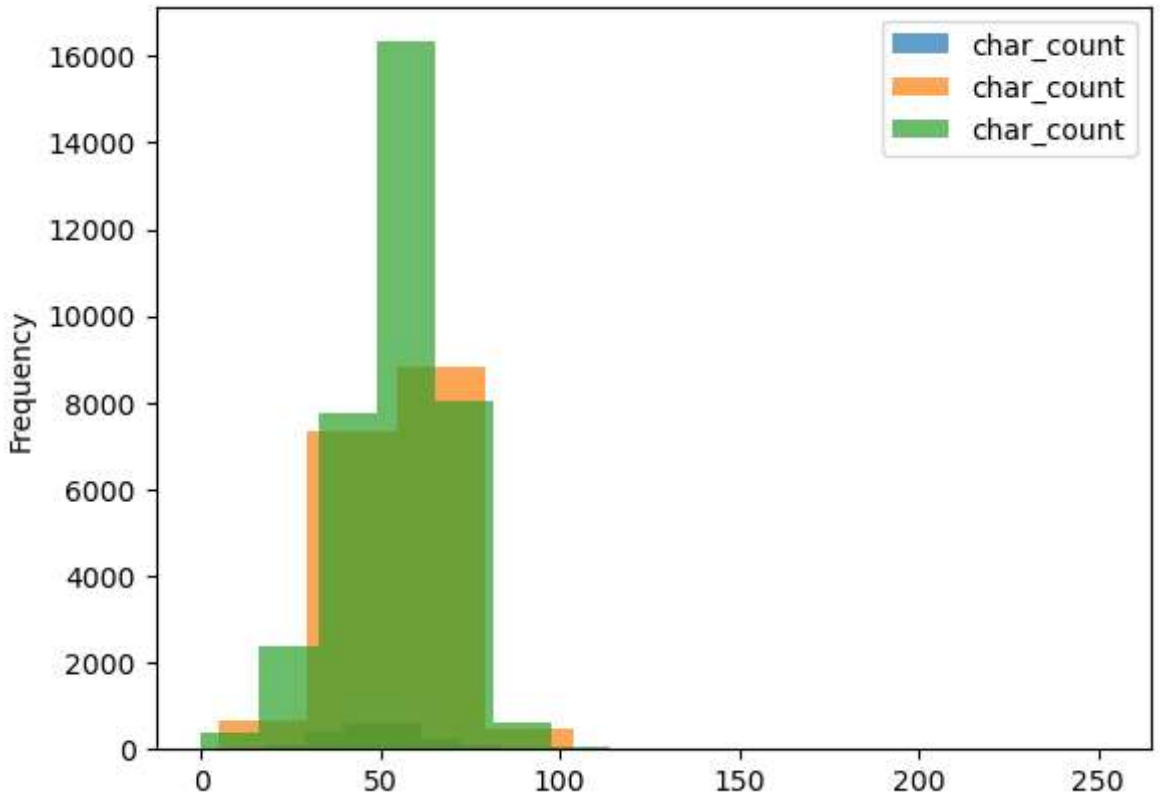
In [48]:
```
dtf['word_count'] = dtf["text"].apply(lambda x: len(str(x).split(" ")))

dtf['char_count'] = dtf["text"].apply(lambda x: sum(len(word) for word in str(

dtf['sentence_count'] = dtf["text"].apply(lambda x: len(str(x).split(".")))

dtf['avg_word_length'] = dtf['char_count'] / dtf['word_count']

dtf['avg_sentence_lenght'] = dtf['word_count'] / dtf['sentence_count']
dtf.head()
```

Out[48]:

	text	y	text_clean	word_count	char_count	sentence_count	avg_word_
13	Twitch Bans Gambling Sites After Streamer Scam...	TECH	twitch ban gambling site streamer scam folk 20...	11	59	1	5.
20	Golden Globes Returning To NBC In January Afte...	ENTERTAINMENT	golden globe returning nbc january year offair	10	51	1	5.
21	Biden Says U.S. Forces Would Defend Taiwan If ...	POLITICS	biden say u force would defend taiwan china in...	10	50	3	5.
24	'Beautiful And Sad At The Same Time': Ukrainia...	POLITICS	beautiful sad time ukrainian cultural festival...	17	85	1	5.
28	James Cameron Says He 'Clashed' With Studio Be...	ENTERTAINMENT	james cameron say clashed studio avatar release	10	58	1	5.

In [53]:
```
categories = ['TECH', 'ENTERTAINMENT', 'POLITICS']
```

In [67]:
```
for i in categories:
    dtf[dtf['y'] == i]['char_count'].plot(kind = 'hist',
                                          alpha = 0.7)
    plt.legend()
```



In [75]:
```
dtf["sentiment"] = dtf['text_clean'].apply(lambda x: TextBlob(x).sentiment.pol
dtf.head()
```

Out[75]:

	text	y	text_clean	word_count	char_count	sentence_count	avg_word
13	Twitch Bans Gambling Sites After Streamer Scam...	TECH	twitch ban gambling site streamer scam folk 20...	11	59	1	5.
20	Golden Globes Returning To NBC In January Afte...	ENTERTAINMENT	golden globe returning nbc january year offair	10	51	1	5.
21	Biden Says U.S. Forces Would Defend Taiwan If ...	POLITICS	biden say u force would defend taiwan china in...	10	50	3	5.
24	‘Beautiful And Sad At The Same Time’: Ukrainia...	POLITICS	beautiful sad time ukrainian cultural festival...	17	85	1	5.
28	James Cameron Says He 'Clashed' With Studio Be...	ENTERTAINMENT	james cameron say clashed studio avatar release	10	58	1	5.

In [74]:
```
from textblob import TextBlob
```

In [84]:
```
dtf_0 = dtf[dtf['sentiment'] == 0]
```

In [87]:
```
dtf_0['text_clean'][43]
```

Out[87]:
```
'phantom opera close broadway next year'
```

In [88]:
```
dtf_neg = dtf[dtf['sentiment'] == -1]
```

In [89]:
```
dtf_neg.head()
```

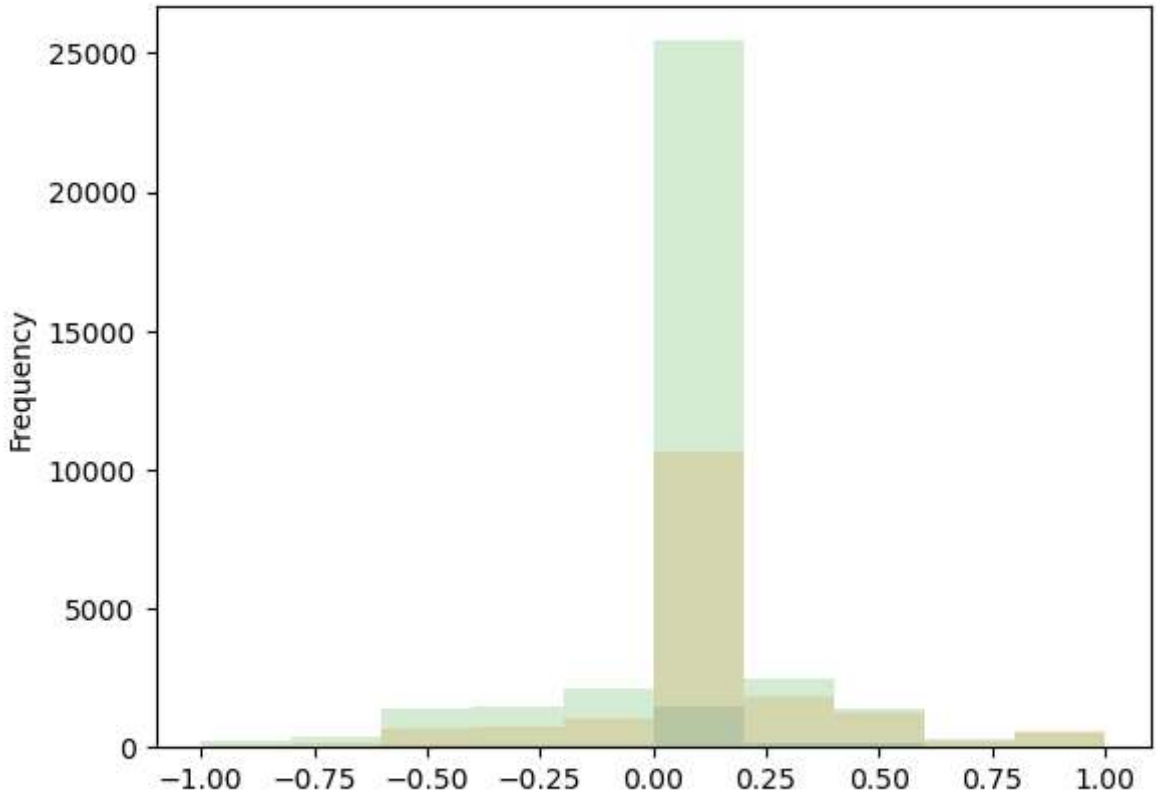
Out[89]:

	text	y	text_clean	word_count	char_count	sentence_count	avg_w
142	Amber Heard's Sister Reacts To 'Disgusting' Jo...	ENTERTAINMENT	amber heards sister reacts disgusting johnny d...	11	59	1	
155	Jared Kushner Blasts 'Nasty Troll' Chrissy Tei...	POLITICS	jared kushner blast nasty troll chrissy teigen...	11	61	1	
364	Ohio Girl's Abortion Doctor Once Targeted In V...	POLITICS	ohio girl abortion doctor targeted vicious kid...	11	68	1	
622	Florence Pugh Calls Out 'Horrible' Reaction To...	ENTERTAINMENT	florence pugh call horrible reaction rumored b...	13	70	1	
1365	Ben Affleck Reveals Why Filming 'Justice Leagu...	ENTERTAINMENT	ben affleck reveals filming justice league wor...	11	65	1	

In [91]:
```
dtf_neg['text_clean'][142]
```

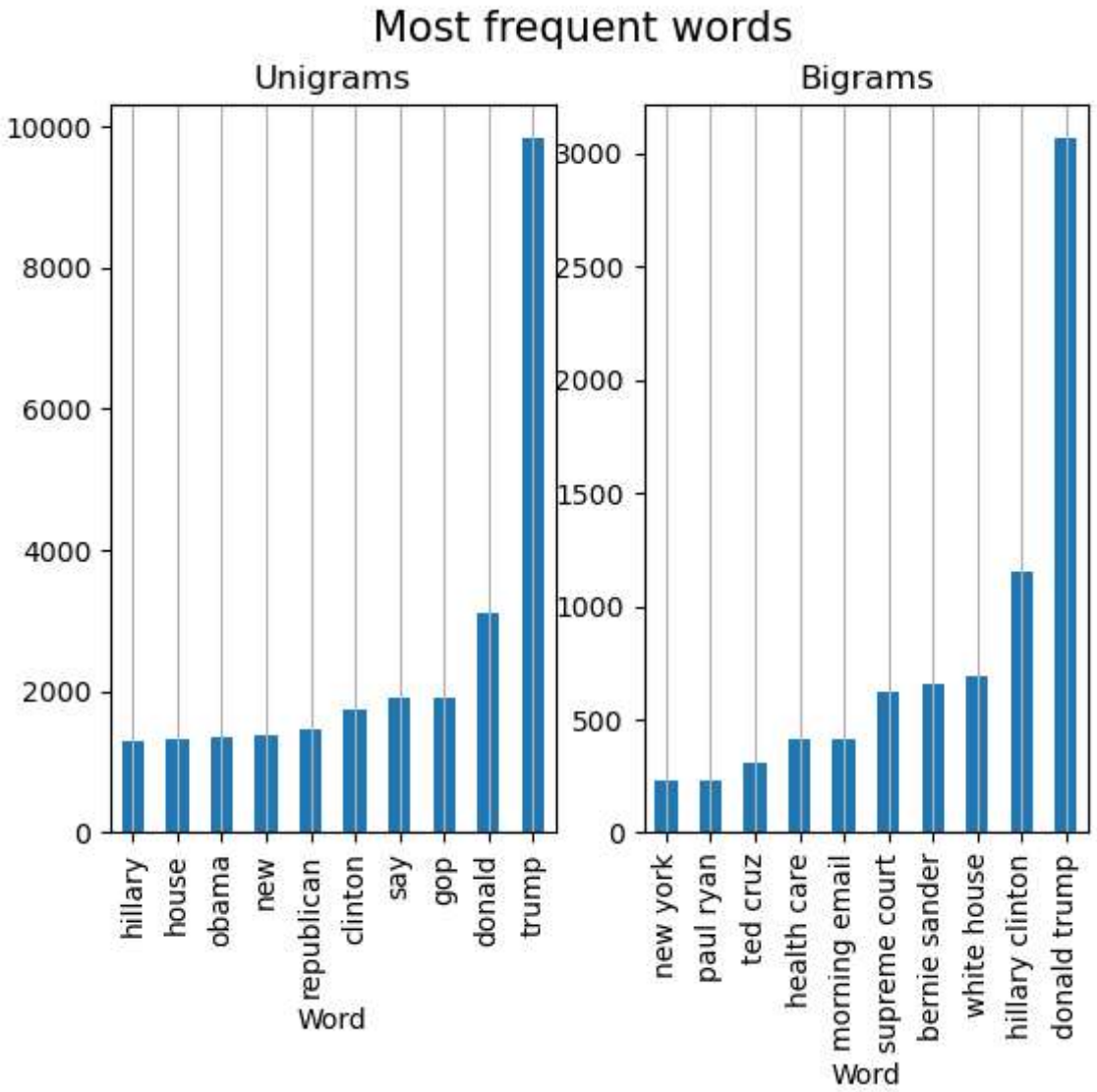
Out[91]:
```
'amber heards sister reacts disgusting johnny depp cameo vmas'
```

In [93]:
```
for i in categories:
    dtf[dtf['y'] == i]['sentiment'].plot(kind = 'hist', alpha = 0.2)
```



In [103]:

```
#n grams
y = "POLITICS"
corpus = dtf[dtf["y"]==y]["text_clean"]
lst_tokens = nltk.tokenize.word_tokenize(corpus.str.cat(sep=" "))
fig, ax = plt.subplots(nrows=1, ncols=2)
plt.figure(figsize = (12, 12))
fig.suptitle("Most frequent words", fontsize=15)

## unigrams
dic_words_freq = nltk.FreqDist(lst_tokens)
dtf_uni = pd.DataFrame(dic_words_freq.most_common(), columns=["Word","Freq"])
dtf_uni.set_index("Word").iloc[:10,:].sort_values(by="Freq").plot(kind="bar",
ax[0].set(ylabel=None)

## bigrams
dic_words_freq = nltk.FreqDist(nltk.ngrams(lst_tokens, 2))
dtf_bi = pd.DataFrame(dic_words_freq.most_common(), columns=["Word","Freq"])
dtf_bi["Word"] = dtf_bi["Word"].apply(lambda x: " ".join(string for string in
dtf_bi.set_index("Word").iloc[:10,:].sort_values(by="Freq").plot(kind="bar", t
ax[1].set(ylabel=None)
plt.show()
```




```
<Figure size 1200x1200 with 0 Axes>
```

To add the most frequently used words as a feature, CountVectorizer from sciki-learn is used.

In [106]:
```
!pip install wordcloud
```

```
Collecting wordcloud
  Downloading wordcloud-1.8.2.2-cp39-cp39-win_amd64.whl (153 kB)
     ------------------------------------ 153.1/153.1 kB 1.3 MB/s eta 0:00:
00
Requirement already satisfied: matplotlib in c:\users\91900\anaconda3\lib\sit
e-packages (from wordcloud) (3.5.2)
Requirement already satisfied: numpy>=1.6.1 in c:\users\91900\anaconda3\lib\s
ite-packages (from wordcloud) (1.21.5)
Requirement already satisfied: pillow in c:\users\91900\anaconda3\lib\site-pa
ckages (from wordcloud) (9.2.0)
Requirement already satisfied: packaging>=20.0 in c:\users\91900\anaconda3\li
b\site-packages (from matplotlib->wordcloud) (21.3)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\91900\anacond
a3\lib\site-packages (from matplotlib->wordcloud) (2.8.2)
Requirement already satisfied: pyparsing>=2.2.1 in c:\users\91900\anaconda3\l
ib\site-packages (from matplotlib->wordcloud) (3.0.9)
Requirement already satisfied: cycler>=0.10 in c:\users\91900\anaconda3\lib\s
ite-packages (from matplotlib->wordcloud) (0.11.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\91900\anaconda3
\lib\site-packages (from matplotlib->wordcloud) (1.4.2)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\91900\anaconda3
\lib\site-packages (from matplotlib->wordcloud) (4.25.0)
Requirement already satisfied: six>=1.5 in c:\users\91900\anaconda3\lib\site-
packages (from python-dateutil>=2.7->matplotlib->wordcloud) (1.16.0)
Installing collected packages: wordcloud
Successfully installed wordcloud-1.8.2.2
```

In [112]:
```
import wordcloud
wc = wordcloud.WordCloud(background_color='white', max_words=1000, max_font_si
wc = wc.generate(str(corpus))
fig = plt.figure(figsize = (10, 6))
plt.axis('off')
plt.imshow(wc, cmap=None)
plt.show()
```

BOW left.

In [ ]: