COVID 19 SENTIMENT ANALYSIS USING VADER AND TEXTBLOB

Mod 9.2 Presentation

Lim Zheng Wei

CONTENTS

INTRO

Objective: See how sentiment varies between countries and over a period of time

Scope: Use VADER and TextBlob to find out sentiment score

We will find articles across the past 2 weeks (23 Aug-5 Sep)

Comparing Singapore, UK and US

1. WEB SCRAPING

https://medium.com/@lzpdatascience/web-scraping-and-text-summarization-of-news-articles-using-python-2ecfb3e71050

Or use Eric's demo on Text Data Collection

```
#Import libraries

from bs4 import BeautifulSoup
import requests
import pandas as pd
```

Article 1: 24 Aug 2020 - https://www.channelnewsasia.com/news/business/pubs-karaoke-nightclubs-covid-19-closing-down-13041928

Lights out, music stops: Still-shuttered pubs, karaoke joints call for help amid COVID-19 pandemic

```
input_file1 = 'C:/Users/zheng/Desktop/Data Science/Presentations/Mod 9.2/SG Articles/SG Articles 1.txt'

with open(input_file1, 'r') as f:
    SGA1 = f.read()
```

```
print(SGA1[:500])
```

```
SINGAPORE: It was early January when PUBking opened its doors to welcome its first customers. But the sight of merry partygoers
clinking their glasses and singing their hearts out lasted for just two months before entertainment venues were forced to shut
as part of COVID-19 control measures. Nearly five months on, the karaoke pub in Outram remains shut as it is excluded from the
list of businesses allowed to reopen. Far from recouping their initial investment of more than S$100,000, its owners h
```

MERGE THEM ALL

```
# Creating a list of filenames
filenames = [input_file1, input_file2, input_file3, input_file4, input_file5, input_file6, input_file7, input_file8, input_file9,

company = ["CNA Article 1", "CNA Article 2", "CNA Article 3", "CNA Article 4", "CNA Article 5", "BBC Article 1", "BBC Article 2",

# Open all 15 in write mode
with open('merged.txt', 'w') as outfile:

    # Iterate through list
    for names in filenames:

        # Open each file in read mode
        with open(names) as infile:

            # read the data from file1 and
            # file2 and write it in file3
            outfile.write(infile.read())

        # Add '\n' to enter data of file2
        # from next line
        outfile.write("\n")
```

```
with open('merged.txt', 'r') as f:
    merge = f.read()
```

CONVERT TO DATAFRAME

	t	e	x	t.1
0	SINGAPORE: It was early January when PUBking o...	NaN	NaN	NaN
1	SINGAPORE: To ensure Singapore's strategic int...	NaN	NaN	NaN
2	SINGAPORE: Polytechnic graduate Andrew Lee was...	NaN	NaN	NaN
3	SINGAPORE: Even though community transmission ...	NaN	NaN	NaN
4	SINGAPORE: The circuit breaker, implemented fr...	NaN	NaN	NaN
5	Anxiety levels among young teenagers dropped d...	NaN	NaN	NaN
6	The number of daily UK cases of coronavirus ha...	NaN	NaN	NaN
7	Universities in the UK are being urged to scra...	NaN	NaN	NaN
8	'I didn't want to send them back' "It's scary ...	NaN	NaN	NaN
9	The government has urged Whitehall bosses to "...	NaN	NaN	NaN
10	As the coronavirus pandemic gained traction in...	NaN	NaN	NaN
11	Americans rank dead last -- by a long way -- a...	NaN	NaN	NaN
12	More than 1,000 students at the University of ...	NaN	NaN	NaN
13	Philadelphia Eagles owner Jeffrey Lurie has cr...	NaN	NaN	NaN
14	More than 410,000 people in the US could die f...	NaN	NaN	NaN

```
df = df.drop(columns=['e','x', 't.1'])
df = df.rename(columns={'t':'Text'})
df['Articles'] = company
```

	Text	Articles
0	SINGAPORE: It was early January when PUBking o...	CNA Article 1
1	SINGAPORE: To ensure Singapore's strategic int...	CNA Article 2
2	SINGAPORE: Polytechnic graduate Andrew Lee was...	CNA Article 3
3	SINGAPORE: Even though community transmission ...	CNA Article 4
4	SINGAPORE: The circuit breaker, implemented fr...	CNA Article 5
5	Anxiety levels among young teenagers dropped d...	BBC Article 1
6	The number of daily UK cases of coronavirus ha...	BBC Article 2
7	Universities in the UK are being urged to scra...	BBC Article 3
8	'I didn't want to send them back' "It's scary ...	BBC Article 4
9	The government has urged Whitehall bosses to "...	BBC Article 5
10	As the coronavirus pandemic gained traction in...	CNN Article 1
11	Americans rank dead last -- by a long way -- a...	CNN Article 2
12	More than 1,000 students at the University of ...	CNN Article 3
13	Philadelphia Eagles owner Jeffrey Lurie has cr...	CNN Article 4
14	More than 410,000 people in the US could die f...	CNN Article 5

```
cols = list(df)
cols[1], cols[0] = cols[0], cols[1]
cols
```

```
['Articles', 'Text']
```

```
df = df.loc[:,cols]

df
```

	Articles	Text
0	CNA Article 1	SINGAPORE: It was early January when PUBking o...
1	CNA Article 2	SINGAPORE: To ensure Singapore's strategic int...
2	CNA Article 3	SINGAPORE: Polytechnic graduate Andrew Lee was...
3	CNA Article 4	SINGAPORE: Even though community transmission ...
4	CNA Article 5	SINGAPORE: The circuit breaker, implemented fr...
5	BBC Article 1	Anxiety levels among young teenagers dropped d...
6	BBC Article 2	The number of daily UK cases of coronavirus ha...
7	BBC Article 3	Universities in the UK are being urged to scra...
8	BBC Article 4	'I didn't want to send them back' "It's scary ...
9	BBC Article 5	The government has urged Whitehall bosses to "...
10	CNN Article 1	As the coronavirus pandemic gained traction in...
11	CNN Article 2	Americans rank dead last -- by a long way -- a...
12	CNN Article 3	More than 1,000 students at the University of ...
13	CNN Article 4	Philadelphia Eagles owner Jeffrey Lurie has cr...
14	CNN Article 5	More than 410,000 people in the US could die f...

2. CLEAN TEXT

```
import string
import spacy
from spacy.lang.en.stop_words import STOP_WORDS
from spacy.lang.en import English

# Create our list of punctuation marks
punctuations = string.punctuation

# Create our list of stopwords
nlp = spacy.load("en_core_web_sm")
stop_words = spacy.lang.en.stop_words.STOP_WORDS

# Load English tokenizer, tagger, parser, NER and word vectors
parser = English()

# Creating our tokenizer function
def spacy_tokenizer(sentence):
    # Creating our token object, which is used to create documents with linguistic annotations.
    mytokens = parser(sentence)

    # Lemmatizing each token and converting each token into lowercase
    mytokens = [ word.lemma_.lower().strip() if word.lemma_ != "-PRON-" else word.lower_ for word in mytokens ]

    # Removing stop words
    mytokens = [ word for word in mytokens if word not in stop_words and word not in punctuations ]

    # return preprocessed list of tokens
    return mytokens
```

df['clea

df['clea

	Articles	Text	clean
1	CNA Article 1	SINGAPORE: It was early January when PUBking o...	singapore early january pubking opened doors w...
2	CNA Article 2	SINGAPORE: To ensure Singapore's strategic int...	singapore ensure singapore strategic interests...
3	CNA Article 3	SINGAPORE: Polytechnic graduate Andrew Lee was...	singapore polytechnic graduate andrew lee init...
4	CNA Article 4	SINGAPORE: Even though community transmission ...	singapore community transmission covid-19 low ...
5	CNA Article 5	SINGAPORE: The circuit breaker, implemented fr...	singapore circuit breaker implemented april ju...
6	BBC Article 1	Anxiety levels among young teenagers dropped d...	anxiety levels young teenagers dropped coronav...
7	BBC Article 2	The number of daily UK cases of coronavirus ha...	number daily uk cases coronavirus risen 1,522 ...
8	BBC Article 3	Universities in the UK are being urged to scra...	universities uk urged scrap plans face face te...
9	BBC Article 4	'I didn't want to send them back' "It's scary ...	want send scary choice says iram kanwal pay fi...
10	BBC Article 5	The government has urged Whitehall bosses to "...	government urged whitehall bosses quickly staf...
11	CNN Article 1	As the coronavirus pandemic gained traction in...	coronavirus pandemic gained traction united st...
12	CNN Article 2	Americans rank dead last -- by a long way -- a...	americans rank dead -- long way -- citizens do...
13	CNN Article 3	More than 1,000 students at the University of ...	1,000 students university alabama tested posit...
14	CNN Article 4	Philadelphia Eagles owner Jeffrey Lurie has cr...	philadelphia eagles owner jeffrey lurie critic...
15	CNN Article 5	More than 410,000 people in the US could die f...	410,000 people die coronavirus january 1 doubl...

3. NAMED ENTITY RECOGNITION(NER)

NER is a forr

Classify nam

ation

```
SGA1_NER = nlp(SGA1)

entities1 =[(i, i.label_, i.label) for i in SGA1_NER.ents]
entities1
```

```
[(early January, 'DATE', 391),
 (first, 'ORDINAL', 396),
 (just two months, 'DATE', 391),
 (Nearly five months, 'DATE', 391),
 (Outram, 'FAC', 9191306739292312949),
 (the Jobs Support Scheme, 'ORG', 383),
 (the past months, 'DATE', 391),
 (one, 'CARDINAL', 397),
 (Alvin Chua, 'PERSON', 380),
 (June, 'DATE', 391),
 (Singapore, 'GPE', 384),
 (Mr Chua's, 'ORG', 383),
 (Heng Swee Keat, 'PERSON', 380),
 (last week, 'DATE', 391),
 (the Ministry of Trade and Industry, 'ORG', 383),
 (The Singapore Nightlife Business Association, 'ORG', 383),
 (about one-third, 'CARDINAL', 397),
```

FROM SPACY IMPORT DISPLACY

```
USA5_NER = nlp(USA5)

displacy.render(USA5_NER, style = "ent",jupyter = True)
```

death rate could reach nearly 3,000 CARDINAL a day by December DATE , an unprecedented number, due in part to "declining vigilance of the public," the IHME ORG expects. For now, the model points to declining mask use in some regions from peak usage in early August DATE . The IHME ORG model is more aggressive in its predictions than others. It comes a day DATE after a new CDC ORG ensemble forecast predicted 211,000 CARDINAL US GPE deaths from Covid-19 by September 26 DATE . Coronavirus PERSON has infected over 6.1 million CARDINAL people nationwide, and more than 190,000 CARDINAL have died, according to Johns Hopkins University ORG . Fauci: US GPE has to get the baseline of cases down Dr. Anthony Fauci PERSON says there is only one way to prevent the death toll reaching the numbers predicted in this new model. "We've got to get our baseline back down to a much lower level," Fauci, the director of the National Institute of Allergy and Infectious Diseases ORG , said on CNN ORG . Currently, the US GPE is seeing about 40,000 CARDINAL cases a day, but if the baseline of cases is lowered, the country could get a better handle on stopping the spread, according to Fauci ORG . And the use of masks would help the country prevent the "scary"

4A. SENTIMENT ANALYSIS - VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a combination of A sentiment lexicon is a list of lexical features (e.g., words) which are generally labelled according to their semantic orientation as either positive or negative.

WHY USE VADER?

1. It works exceedingly well on social media type text, yet readily generalizes to multiple domains
2. It doesn't require any training data but is constructed from a generalizable, valence-based, human-curated gold standard sentiment lexicon
3. It is fast enough to be used online with streaming data, and
4. It does not severely suffer from a speed-performance tradeoff.

```python
#!pip install vaderSentiment
```

```python
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
analyzer = SentimentIntensityAnalyzer()
```

```python
#Add VADER metrics to dataframe
df['compound_text'] = [analyzer.polarity_scores(v)['compound'] for v in df['Text']]
df['neg_text'] = [analyzer.polarity_scores(v)['neg'] for v in df['Text']]
df['neu_text'] = [analyzer.polarity_scores(v)['neu'] for v in df['Text']]
df['pos_text'] = [analyzer.polarity_scores(v)['pos'] for v in df['Text']]
```

```python
#Add VADER metrics to dataframe
df['compound_clean'] = [analyzer.polarity_scores(v)['compound'] for v in df['clean']]
df['neg_clean'] = [analyzer.polarity_scores(v)['neg'] for v in df['clean']]
df['neu_clean'] = [analyzer.polarity_scores(v)['neu'] for v in df['clean']]
df['pos_clean'] = [analyzer.polarity_scores(v)['pos'] for v in df['clean']]
```

SCORING SYSTEM

The Positive, Negative and Neutral scores represent the proportion of text that falls in these categories. This means our sentence was rated as 11.1% Positive, 84.1% Neutral and 4.9% Negative. Hence all these should add up to 1.

The Compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive).

	Articles	Text	clean	compound_text	neg_text	neu_text	pos_text	compound_clean	neg_clean	neu_clean	pos_clean
1	CNA Article 1	SINGAPORE: It was early January when PUBking o...	singapore early january pubking opened doors w...	0.9984	0.049	0.841	0.111	0.9973	0.067	0.759	0.174

VADER SCORING METRICS

1. Punctu
magnitud
orientatic nse
than “The f (!),
increases

```
#Baseline sentence
sentiment_analyzer_scores('The food here is good')
```

```
The food here is good------------------- {'neg': 0.0, 'neu': 0.58, 'pos': 0.42,
'compound': 0.4404}
```

```
#Punctuation
print(sentiment_analyzer_scores('The food here is good!'))
print(sentiment_analyzer_scores('The food here is good!!'))
print(sentiment_analyzer_scores('The food here is good!!!'))
```

```
The food here is good!------------------ {'neg': 0.0, 'neu': 0.556, 'pos': 0.44
4, 'compound': 0.4926}
None
The food here is good!!----------------- {'neg': 0.0, 'neu': 0.534, 'pos': 0.46
6, 'compound': 0.5399}
None
The food here is good!!!---------------- {'neg': 0.0, 'neu': 0.514, 'pos': 0.48
6, 'compound': 0.5826}
None
```

2. Capit

sentimen ... ed

words, ir

example ... an

“The foo

```
#Baseline sentence
sentiment_analyzer_scores('The food here is great!')
```

```
The food here is great!----------------- {'neg': 0.0, 'neu': 0.477, 'pos': 0.52
3, 'compound': 0.6588}
```

```
#Capitalisation
sentiment_analyzer_scores('The food here is GREAT!')
```

```
The food here is GREAT!----------------- {'neg': 0.0, 'neu': 0.438, 'pos': 0.56
2, 'compound': 0.729}
```

3. Degre
sentimen ... sity.
For exar ... ense
than “Th
margina

```
#Baseline sentence
sentiment_analyzer_scores('The service here is good')
```

```
The service here is good---------------- {'neg': 0.0, 'neu': 0.58, 'pos': 0.42,
'compound': 0.4404}
```

```
#Degree Modifiers
print(sentiment_analyzer_scores('The service here is extremely good'))
print(sentiment_analyzer_scores('The service here is marginally good'))
```

```
The service here is extremely good------ {'neg': 0.0, 'neu': 0.61, 'pos': 0.39,
'compound': 0.4927}
None
The service here is marginally good----- {'neg': 0.0, 'neu': 0.657, 'pos': 0.34
3, 'compound': 0.3832}
None
```

4. Conjunctions: Use of conjunctions like “but” signals a shift in sentiment polarity, with the sentiment of the text following the conjunction being dominant. “The food here is great, but the service is horrible” has mixed sentiment, with the latter half dictating the overall rating.

```
#Conjunctions
sentiment_analyzer_scores('The food here is great, but the service is horrible')
```

```
The food here is great, but the service is horrible {'neg': 0.31, 'neu': 0.523,
'pos': 0.167, 'compound': -0.4939}
```

5. Preceding Tri-gram: By examining the tri-gram preceding a sentiment-laden lexical feature, we catch nearly 90% of cases where negation flips the polarity of the text. A negated sentence would be “The food here isn’t really all that great”.

4B. SENTIMENT ANALYSIS - TEXTBLOB

Returns in the form of Sentiment(polarity, subjectivity)

Polarity score : Range between -1.0 and 1.0

Subjectivity score : Range between 0.0 and 1.0 where 0.0 is very objective and 1.0 is very subjective.

```
from textblob import TextBlob
```

```
#load the descriptions into textblob
desc_blob = [TextBlob(desc) for desc in df['Text']]
#add the sentiment metrics to the dataframe
df['tb_Pol_text'] = [b.sentiment.polarity for b in desc_blob]
df['tb_Subj_text'] = [b.sentiment.subjectivity for b in desc_blob]
```

```
#load the descriptions into textblob
desc_blob = [TextBlob(desc) for desc in df['clean']]
#add the sentiment metrics to the dataframe
df['tb_Pol_clean'] = [b.sentiment.polarity for b in desc_blob]
df['tb_Subj_clean'] = [b.sentiment.subjectivity for b in desc_blob]
```

5. VISUALISATION

A. Compare a country over a 2 week span

B. Compare the 3 countries at a specific period of time

5A

```
SG = df[0:5]
SG
```

	Articles	Text	clean	compound_text	neg_text	neu_text	pos_text	compound_clean	neg_clean	neu_clean	pos_clean	tb_Pol_text
1	CNA Article 1	SINGAPORE: It was early January when PUBking o...	singapore early january pubking opened doors w...	0.9984	0.049	0.841	0.111	0.9973	0.067	0.759	0.174	0.054668
2	CNA Article 2	SINGAPORE: To ensure Singapore's strategic int...	singapore ensure singapore strategic interests...	0.9614	0.077	0.795	0.127	0.9670	0.116	0.682	0.202	0.085526
3	CNA Article 3	SINGAPORE: Polytechnic graduate Andrew Lee was...	singapore polytechnic graduate andrew lee init...	0.9998	0.032	0.870	0.098	0.9996	0.047	0.803	0.150	0.082470
4	CNA Article 4	SINGAPORE: Even though community transmission ...	singapore community transmission covid-19 low ...	0.9979	0.020	0.879	0.101	0.9959	0.030	0.815	0.155	0.129182
5	CNA Article 5	SINGAPORE: The circuit breaker, implemented fr...	singapore circuit breaker implemented april ju...	0.9965	0.040	0.867	0.093	0.9950	0.054	0.806	0.140	0.128964

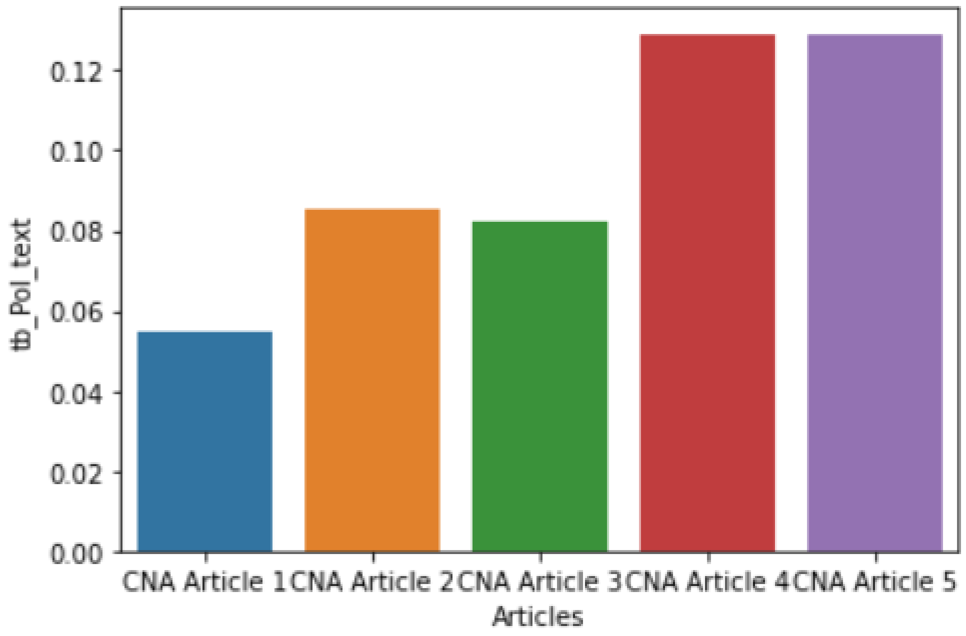
SG - OVERALL SENTIMENT

```
sns.barplot(x='Articles', y='compound_text', data = SG)
```

<AxesSubplot:xlabel='Articles', ylabel='compound_text'>


```
sns.barplot(x='Articles', y='tb_Pol_text', data = SG)
```

<AxesSubplot:xlabel='Articles', ylabel='tb_Pol_text'>



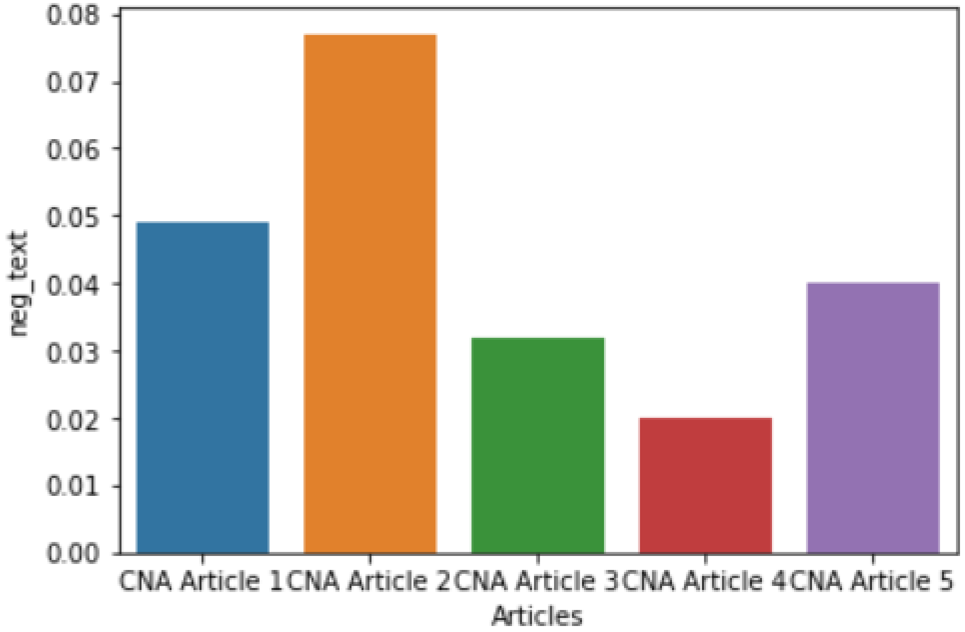
BREAKDOWN

```
sns.barplot(x='Articles', y='pos_text', data = SG)
```

```
<AxesSubplot:xlabel='Articles', ylabel='pos_text'>
```


```
sns.barplot(x='Articles', y='neg_text', data = SG)
```

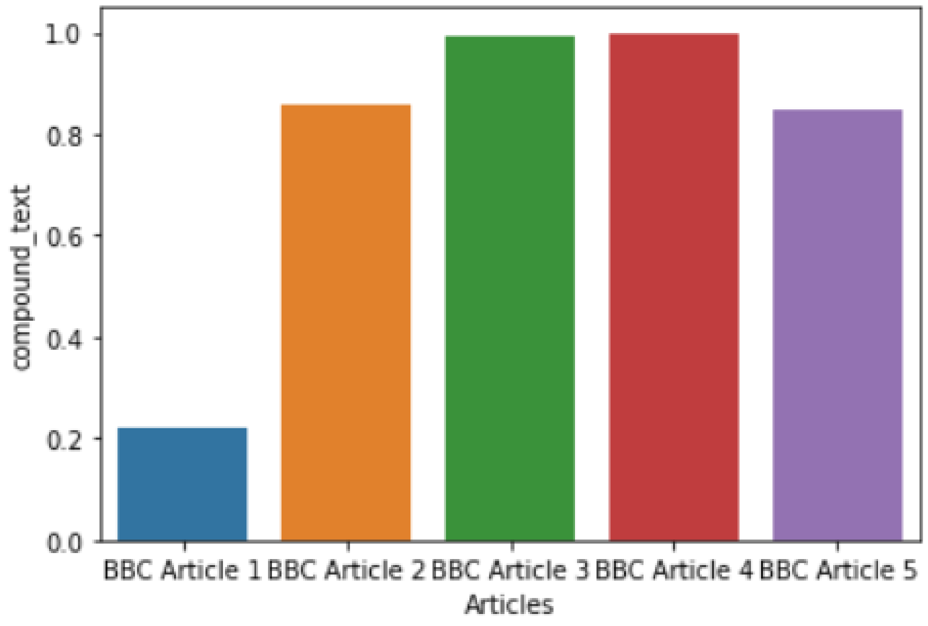
```
<AxesSubplot:xlabel='Articles', ylabel='neg_text'>
```



UK – OVERALL SENTIMENT

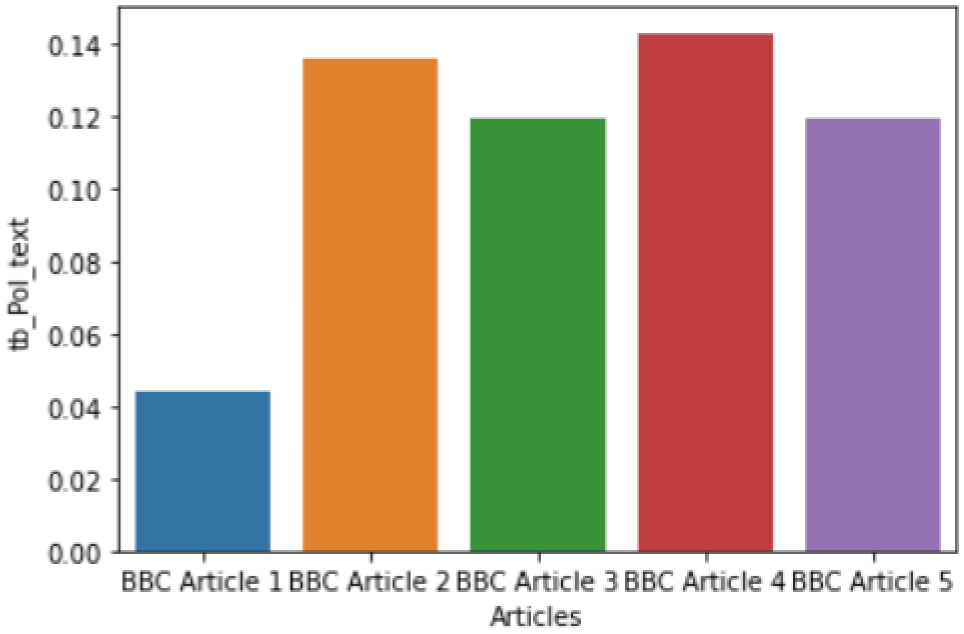
```
sns.barplot(x='Articles', y='compound_text', data = UK)
```

<AxesSubplot:xlabel='Articles', ylabel='compound_text'>




```
sns.barplot(x='Articles', y='tb_Pol_text', data = UK)
```

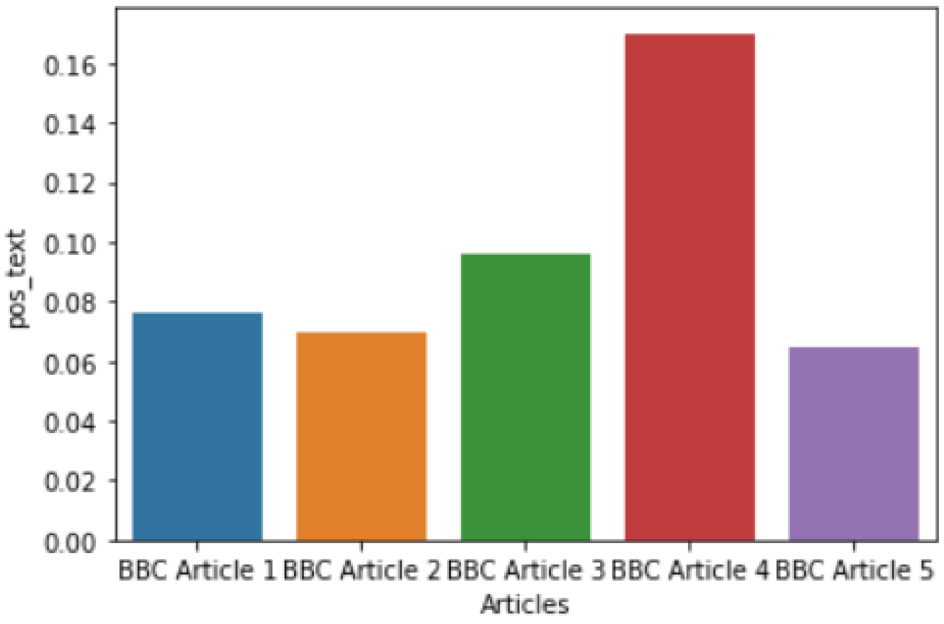
<AxesSubplot:xlabel='Articles', ylabel='tb_Pol_text'>



BREAKDOWN

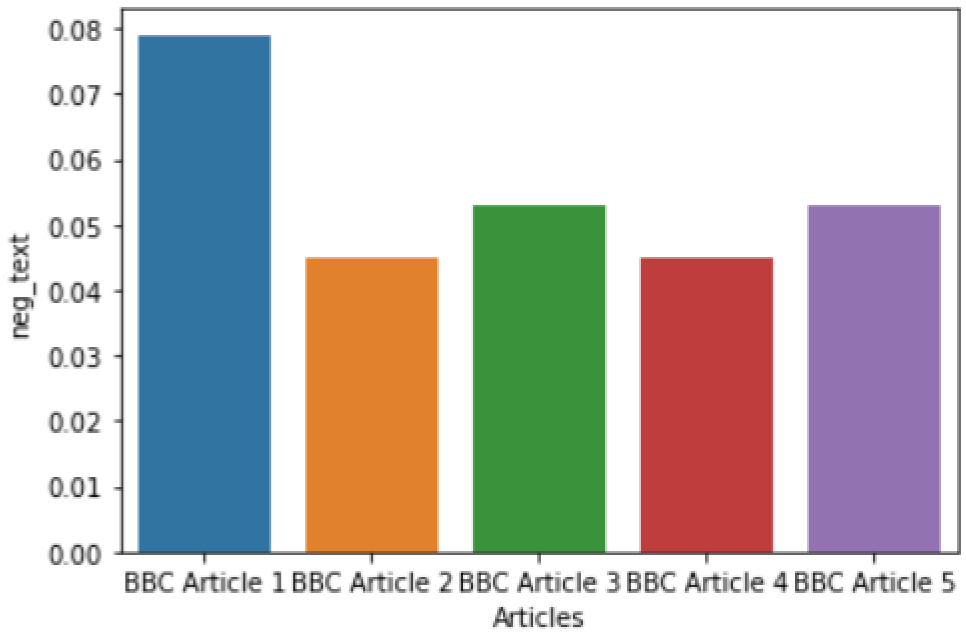
```
sns.barplot(x='Articles', y='pos_text', data = UK)
```

```
<AxesSubplot:xlabel='Articles', ylabel='pos_text'>
```




```
sns.barplot(x='Articles', y='neg_text', data = UK)
```

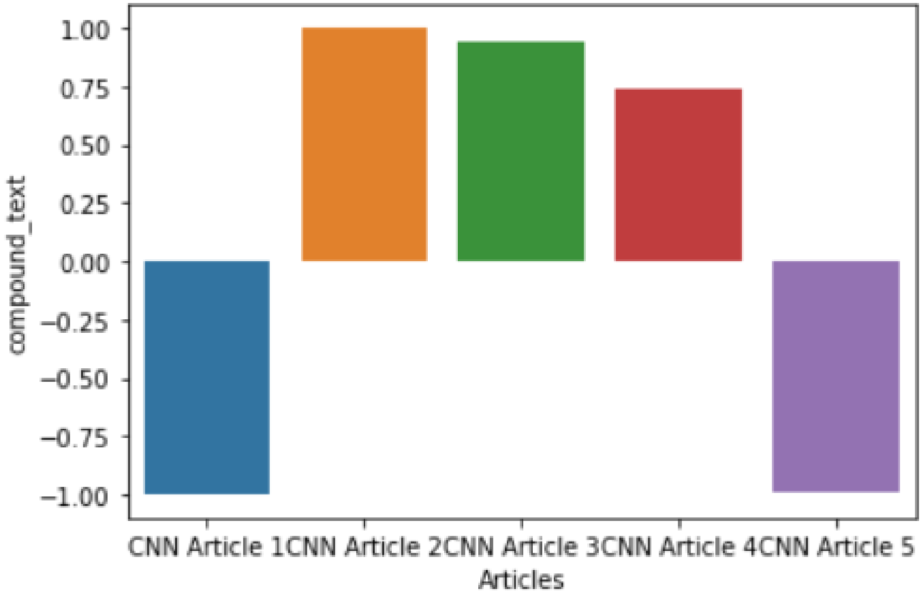
```
<AxesSubplot:xlabel='Articles', ylabel='neg_text'>
```



USA – OVERALL SENTIMENT

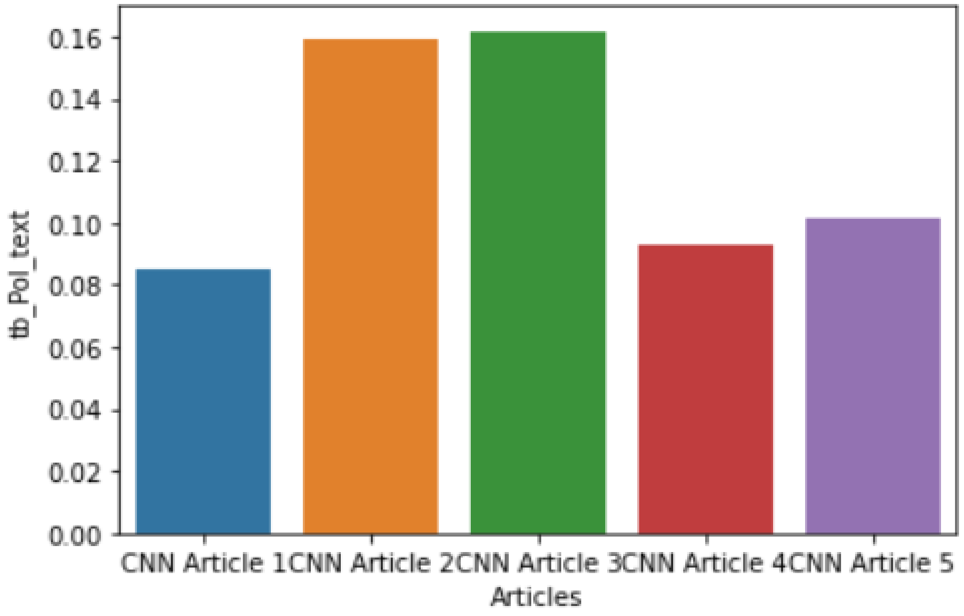
```
sns.barplot(x='Articles', y='compound_text', data = US)
```

```
<AxesSubplot:xlabel='Articles', ylabel='compound_text'>
```




```
sns.barplot(x='Articles', y='tb_Pol_text', data = US)
```

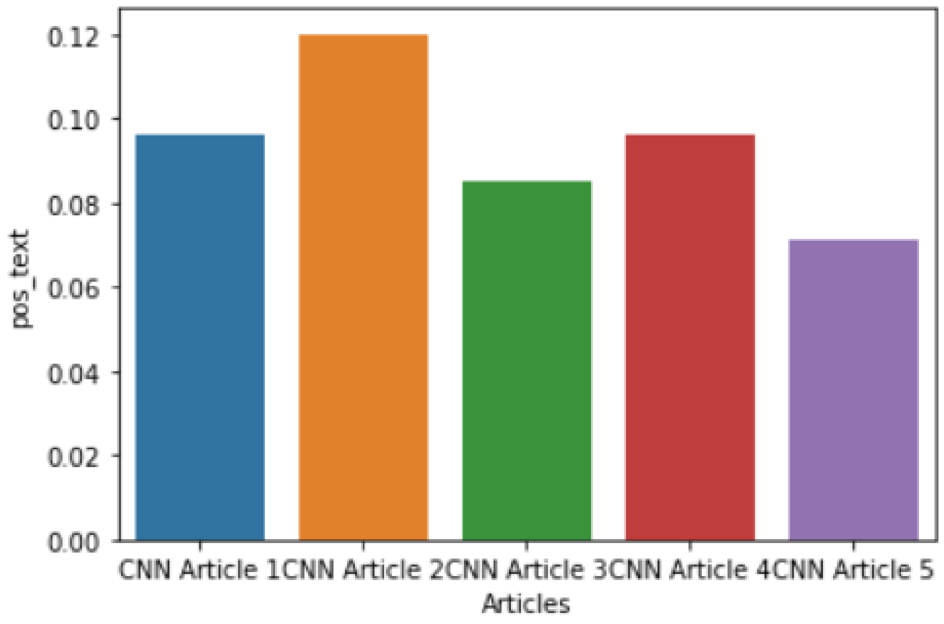
```
<AxesSubplot:xlabel='Articles', ylabel='tb_Pol_text'>
```



BREAKDOWN

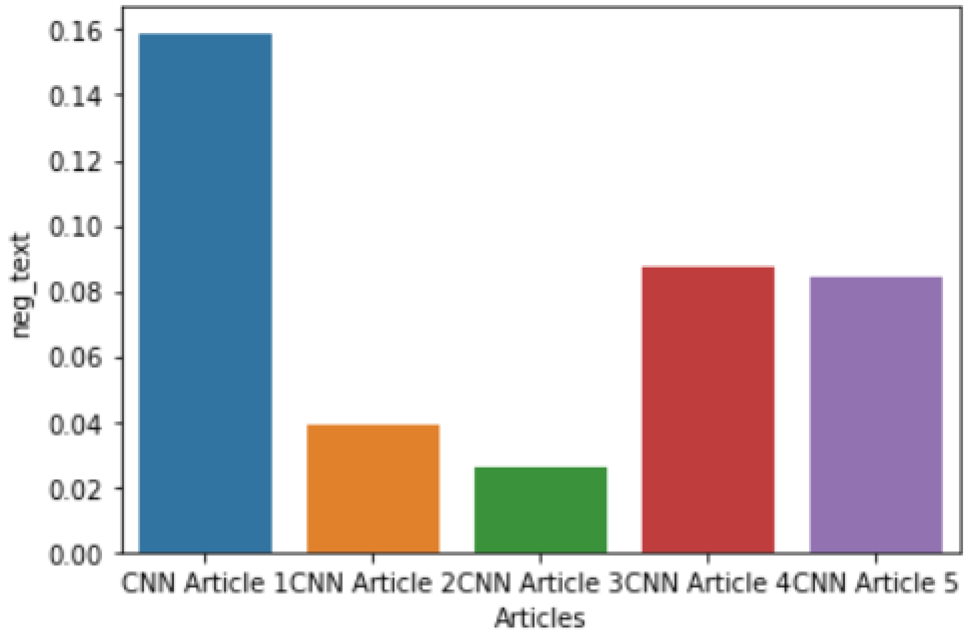
```
sns.barplot(x='Articles', y='pos_text', data = US)
```

```
<AxesSubplot:xlabel='Articles', ylabel='pos_text'>
```




```
sns.barplot(x='Articles', y='neg_text', data = US)
```

```
<AxesSubplot:xlabel='Articles', ylabel='neg_text'>
```

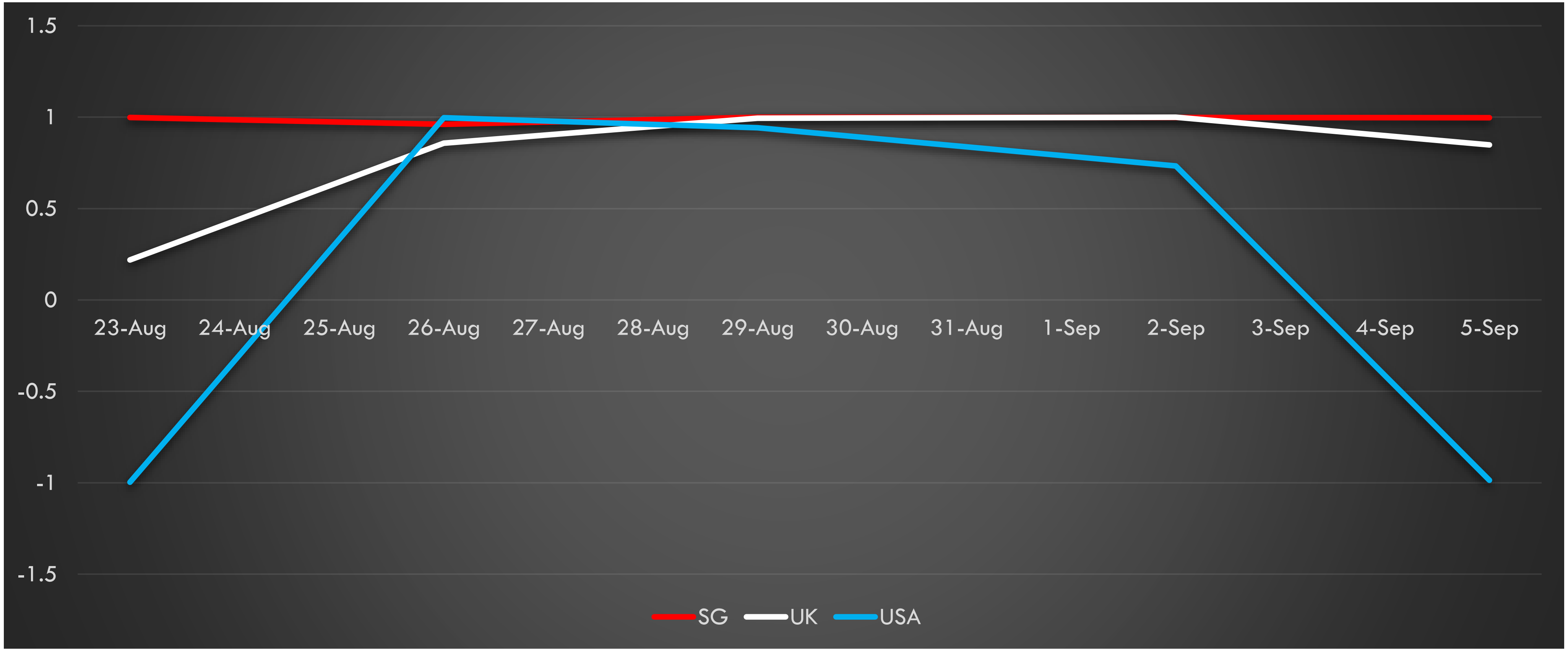


5B

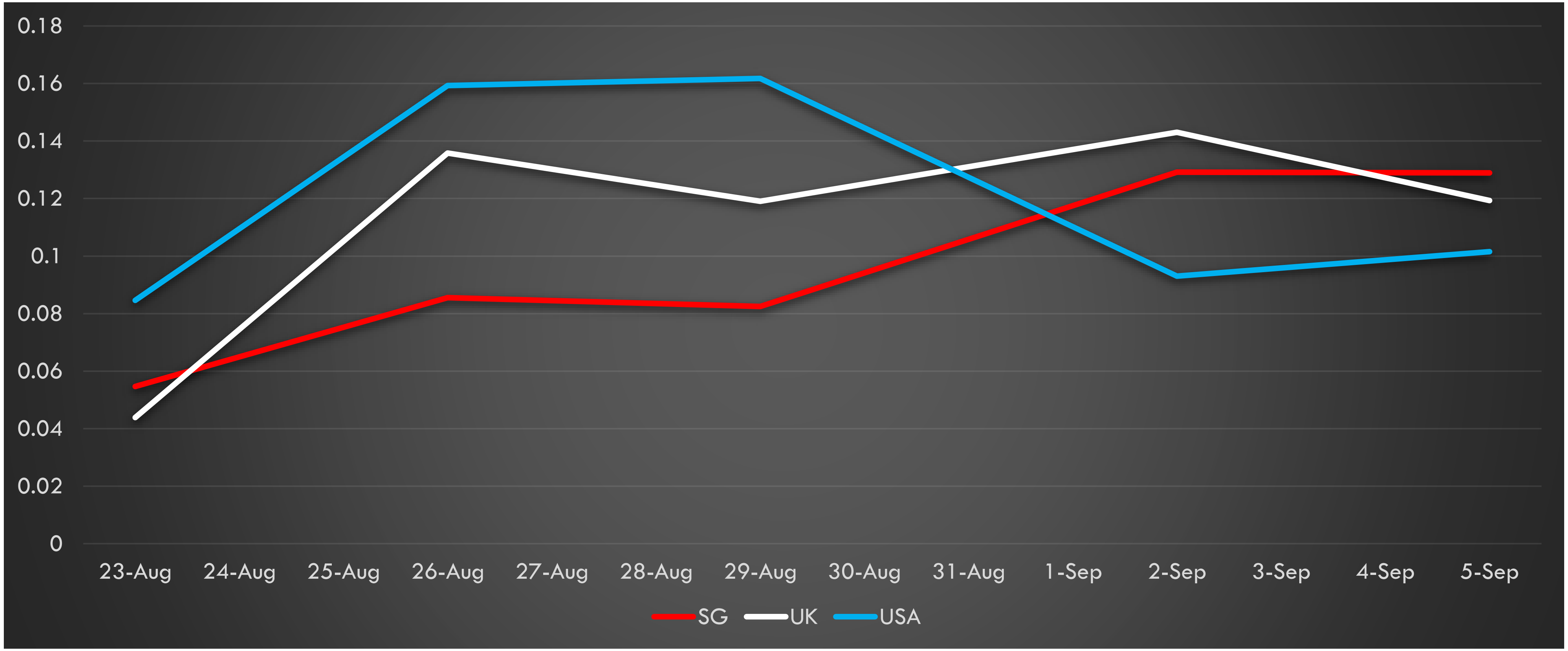
```
TL1 = df.loc[ (df['Articles'] == 'CNA Article 1') | (df['Articles'] == 'BBC Article 1') | (df['Articles'] == 'CNN Article 1') ]
TL1
```

	Articles	Text	clean	compound_text	neg_text	neu_text	pos_text	compound_clean	neg_clean	neu_clean	pos_clean	tb_Pol_text	tb_Subj_
1	CNA Article 1	SINGAPORE: It was early January when PUBking o...	singapore early january pubking opened doors w...	0.9984	0.049	0.841	0.111	0.9973	0.067	0.759	0.174	0.054668	0.442
6	BBC Article 1	Anxiety levels among young teenagers dropped d...	anxiety levels young teenagers dropped coronav...	0.2187	0.079	0.845	0.076	0.0772	0.133	0.738	0.129	0.043864	0.391
11	CNN Article 1	As the coronavirus pandemic gained traction in...	coronavirus pandemic gained traction united st...	-0.9976	0.159	0.746	0.096	-0.9982	0.256	0.598	0.146	0.084591	0.439

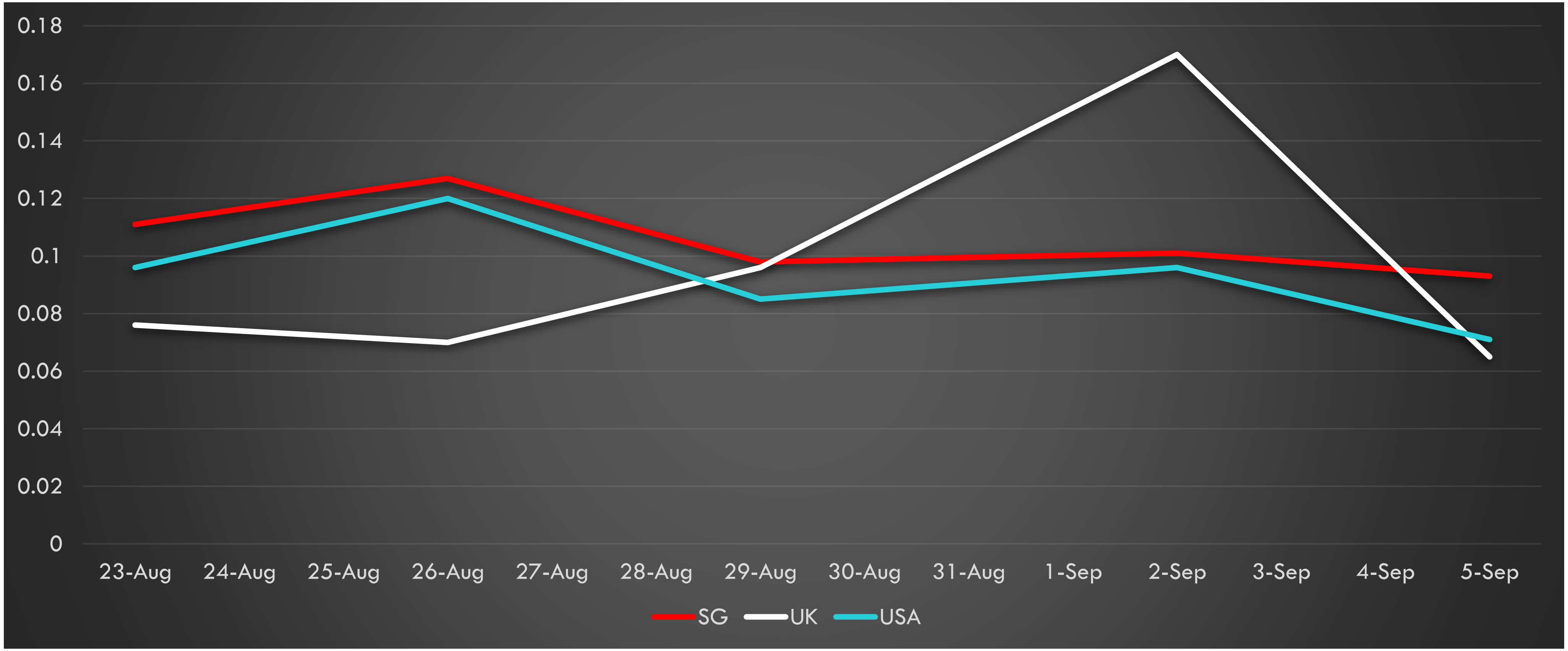
OVERALL SENTIMENT



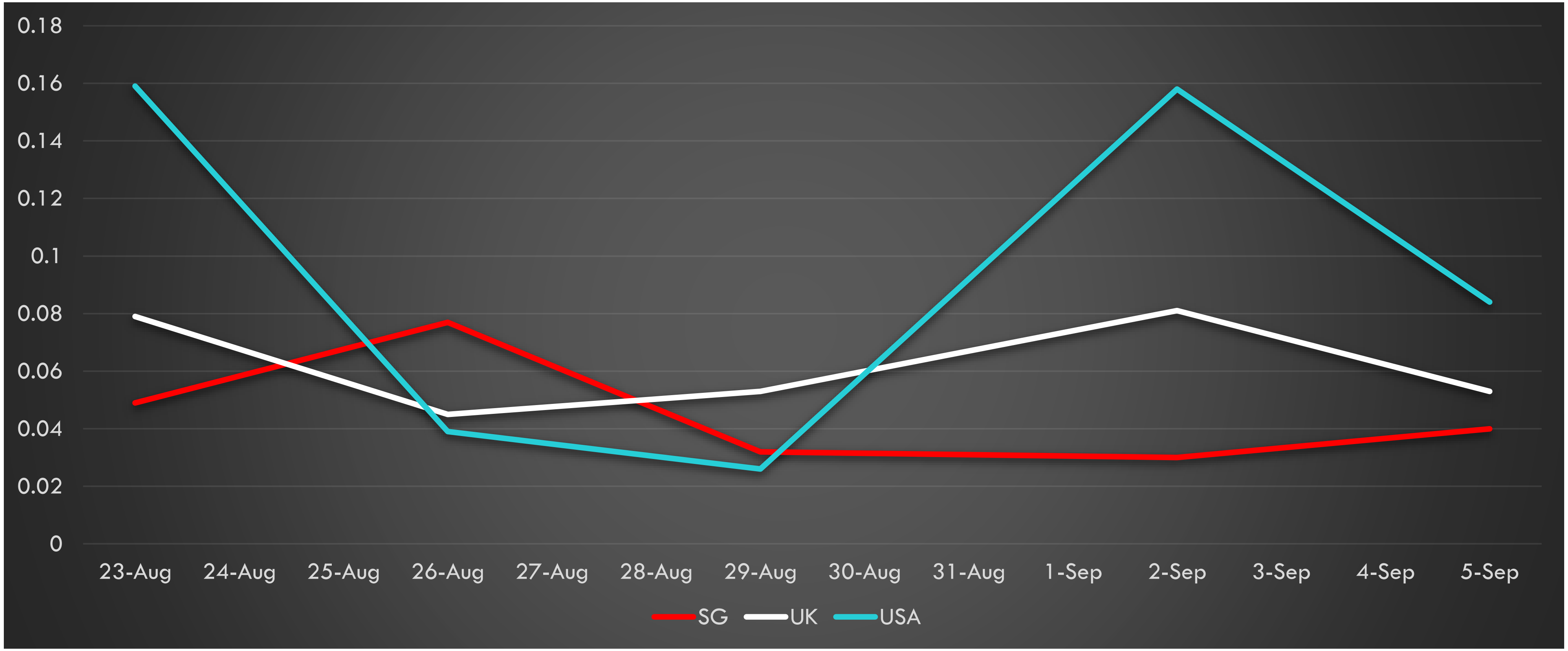
POLARITY SCORE



POSITIVE TEXT



NEGATIVE TEXT



OBSERVATION

No need to clean text for sentiment analysis

Excel can compliment with Python

Can consider doing 5B in Time Series

CONCLUSION

All 3 countries are slightly feeling positive about covid-19

US has spells feeling negative

Shows importance of good governance, how the public feels about the measures

The impacts of covid-19 are real

Take care everybody

Stay safe, wear a mask when going outdoors

REFERENCES

https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f