Sentiment Analysis of Covid Vaccine Tweets

By: Nikita Case

Introduction

- Social Media is no longer just a way to catch up on cat videos
- 42% of Canadians get their news through social media
- It has been shown to influence public opinion and even politics
- So what do twitter users say about COVID Vaccines?

Goals

Determine whether:

Tweets about the Covid Vaccines are more positive or negative

Positive or negative tweets get more likes and retweets

There is any difference in sentiment between tweets mentioning different vaccine companies

Data Collection

Data was collected from the twitter developer API using the Tweepy module

```
# Tweepy search parameters
search_terms = ['covid+vaccine OR covid-19+vaccine\
                OR pfizer OR pfizer-biontech OR
                        pfizerbiontech\
                OR moderna \
                OR astrazeneca OR astra+zeneca \
                -filter:retweets']

# Tweet collection
tweets = tweepy.Cursor(api.search, max_id = max_id,\
    lang ='en',\
    q=search_terms, tweet_mode='extended').items(limit)
```

	created	id_str	text	screen_name	location	followers_count	user_favourites_count	verified	status_count	geo	coords
0	2021-03-03 23:59:59	1367263516457979910	Shieldk2 Yes! My wife works for Pfizer and has...	Sobres74	Seattle, WA	91	2960	False	3192	NaN	NaN
1	2021-03-03 23:59:59	1367263515656867840	oleary_ray: "I will not be taking the Covid va...	cooksonm7	Auckland, New Zealand	555	25880	False	4078	NaN	NaN
2	2021-03-03 23:59:59	1367263515388567563	CDC's report on J&J COVID-19 vaccine - "AC...	bchaiken	Boston, MA USA	511	4	False	380	NaN	NaN
3	2021-03-03 23:59:58	1367263513257930752	ByYourLogic: the Pfizer vaccine makes your dic...	OldManVEVO	NaN	8	323	False	70	NaN	NaN
4	2021-03-03 23:59:56	1367263503782928384	RobDownenChron: The Archdiocese says the new J...	MagEGordon	Houston, TX	4779	7215	True	27559	NaN	NaN

Sample of data frame with tweet data

Text Cleaning

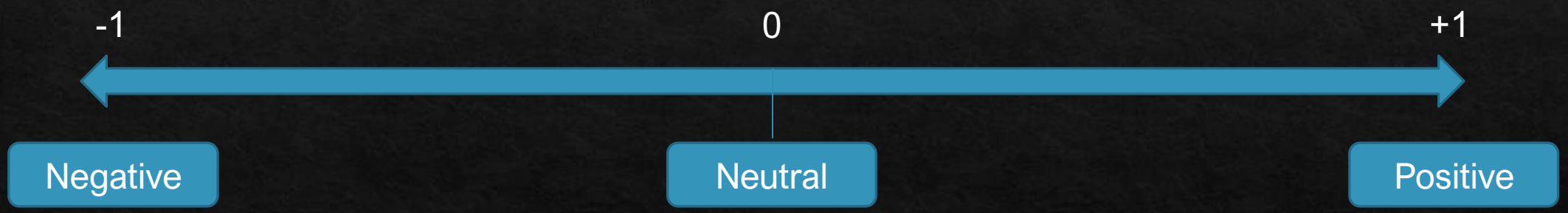
```
# Remove extra rows
tweets_df = tweets_df.drop_duplicates(subset=['id_str'])
tweets_df = tweets_df.dropna(subset=['id_str', 'screen_name'])
```

```
# Makes tweets more readable
def CleanTweets(txt):
    # remove @ from username
    txt = txt.replace('@','')

    # remove RT
    txt = re.sub(r'RT[\s]+','',txt)

    # remove # but leave txt
    txt = re.sub(r'#','', txt)

    # remove hyperlinks
    txt = re.sub(r'https?:\/\/\S+', '', txt)

    txt = re.sub('\n','', txt)
    return txt
```

```
# Clean the text field
tweets_df['text'] = tweets_df['text'].apply(CleanTweets)
```

Polarity

- Polarity is a measure of the emotion of the text



Sentiment Analysis

```python
# to get subjectivity
def RateSubjectivity(txt):
    return TextBlob(txt).sentiment.subjectivity
```

```python
# to get polarity
def RatePolarity(txt):
    return TextBlob(txt).sentiment.polarity
```

```python
# add worded sentiment based on polarity score
def GetSentiment(num):
    if num < 0:
        return 'negative'
    elif num == 0:
        return 'neutral'
    else:
        return 'positive'
```

```python
# Create Columns
def RateTweets(df):
    df['Subjectivity'] = df['text'].apply(RateSubjectivity)
    df['Polarity'] = df['text'].apply(RatePolarity)
    df['Sentiment'] = df['Polarity'].apply(GetSentiment)
    return df
```

Comparing Polarity By Company Mentions

```
# Create groupby object by manufacturer
by_man = polarity_df.groupby(by=['manufacturer'])
```

```
# Perform value counts by polarity
# Normalised because of the large gaps between polarity points
pf = by_man['Polarity'].value_counts(normalize=True).loc['pf']
az = by_man['Polarity'].value_counts(normalize=True).loc['az']
mo = by_man['Polarity'].value_counts(normalize=True).loc['mo']
un = by_man['Polarity'].value_counts(normalize=True).loc['0']
```

```
# Combine polarity counts into datafram and visualize
polarity = pd.concat(
    [pf,az,mo,un],
    axis=1,
    keys=['Pfizer-BioNTech','AstraZeneca','Moderna','Unknown']
)
polarity = polarity.reset_index()
polarity
```

	Polarity	Pfizer-BioNTech	AstraZeneca	Moderna	Unknown
0	-1.0	0.000715	0.000410	0.000517	0.001169
1	-0.9	0.000204	0.000072	0.000222	0.000102
2	-0.8	0.000817	0.000868	0.002661	0.000978
3	-0.7	0.000664	0.000554	0.001922	0.000559
4	-0.6	0.001123	0.004121	0.006430	0.002013
5	-0.5	0.005463	0.002988	0.003769	0.006580
6	-0.4	0.009394	0.009254	0.006652	0.007774
7	-0.3	0.009854	0.011182	0.009534	0.009222
8	-0.2	0.032217	0.028436	0.025868	0.024281
9	-0.1	0.032727	0.097646	0.039394	0.038147
10	0.0	0.415348	0.631419	0.463636	0.402826
11	0.1	0.109415	0.050871	0.134146	0.085509
12	0.2	0.148065	0.057691	0.151367	0.163213
13	0.3	0.065404	0.017254	0.044346	0.051605
14	0.4	0.054529	0.021038	0.041242	0.028766
15	0.5	0.028183	0.044365	0.024316	0.055302
16	0.6	0.066221	0.005808	0.022025	0.033720
17	0.7	0.002757	0.001687	0.003622	0.005577
18	0.8	0.010671	0.001542	0.012121	0.043876
19	0.9	0.001327	0.000361	0.002143	0.000553
20	1.0	0.004901	0.012435	0.004065	0.038229

Manufacturer	Retweets	Likes	Subjectivity	Polarity
Unknown	3663.978164	4.151971	0.339735	0.172315
AstraZeneca	359.724414	2.170494	0.216722	0.043027
Moderna	1486.531633	4.978271	0.291840	0.092520
Pfizer	167.427040	5.882569	0.346579	0.130522

Comparing Polarity Scores

Tweets with AstraZeneca mentions vs Tweets with Moderna mentions

```
az_mo = stats.ttest_ind(az_df.Polarity, mo_df.Polarity, equal_var = False)
az_mo
```

```
Ttest_indResult(statistic=-23.059598150244078, pvalue=2.988406103494073e-116)
```

Tweets with Moderna mentions vs Tweets with Pfizer mentions

```
mo_pf = stats.ttest_ind(mo_df.Polarity, pf_df.Polarity, equal_var = False)
mo_pf
```

```
Ttest_indResult(statistic=-14.884349261786252, pvalue=6.238421682411701e-50)
```

Creating Word Clouds

```
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
```

```
custom_words =['covid', 'covid19', 'covid 19','covid-19','vaccine'\
               ,'covid vaccine','covid 19 vaccine','covid-19 vaccine']
stop_words = stop_words + custom_words
# stop_words
```

```
def Tokenize(txt):
    return TextBlob(txt).words
```

```
def RemoveStopWords(lst):
    no_stopwords = [word for word in lst if not word in stop_words]
    return no_stopwords
```

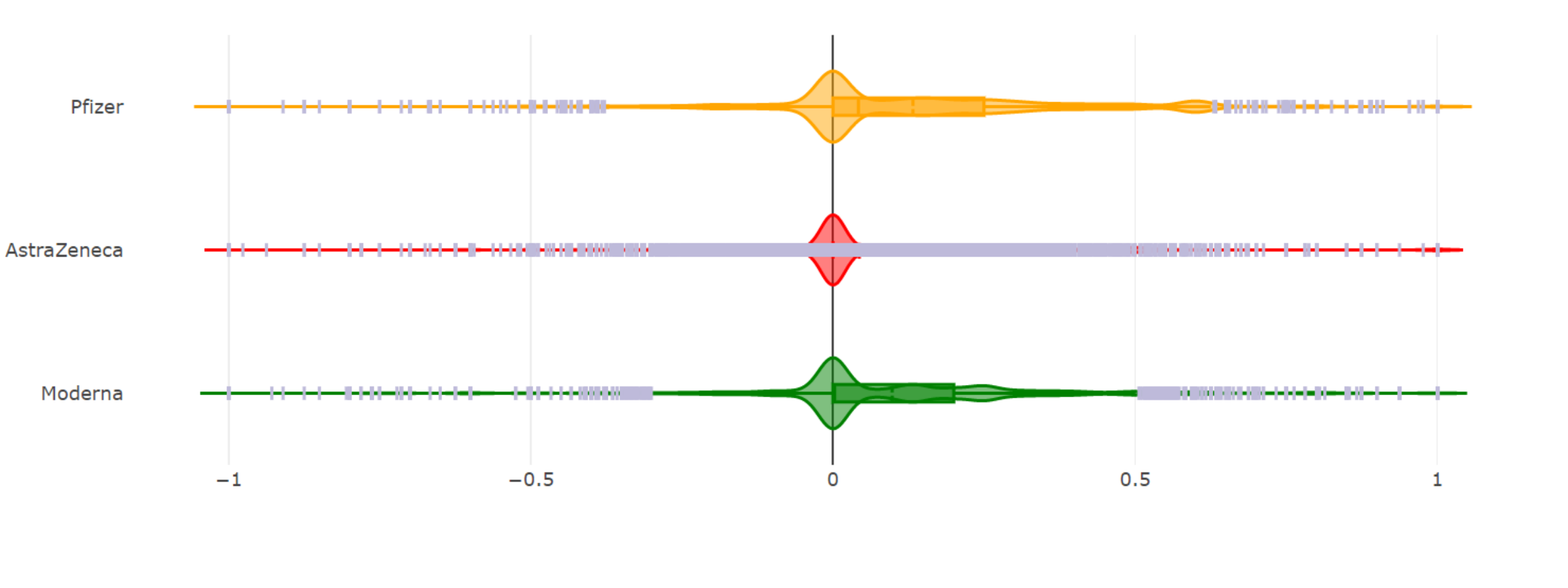
```
def Lemmatize(lst):
    lemmas = [wnl.lemmatize(item) for item in lst]
    return lemmas
```

Visualizations

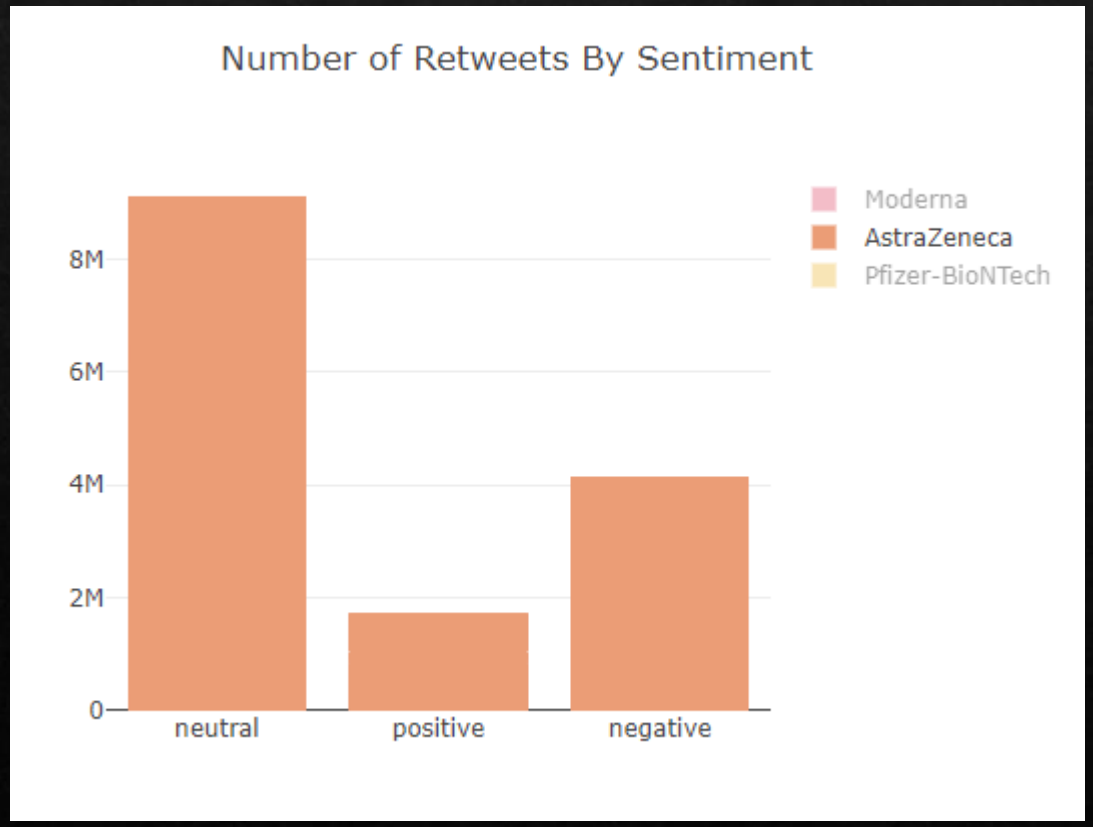
AstraZeneca

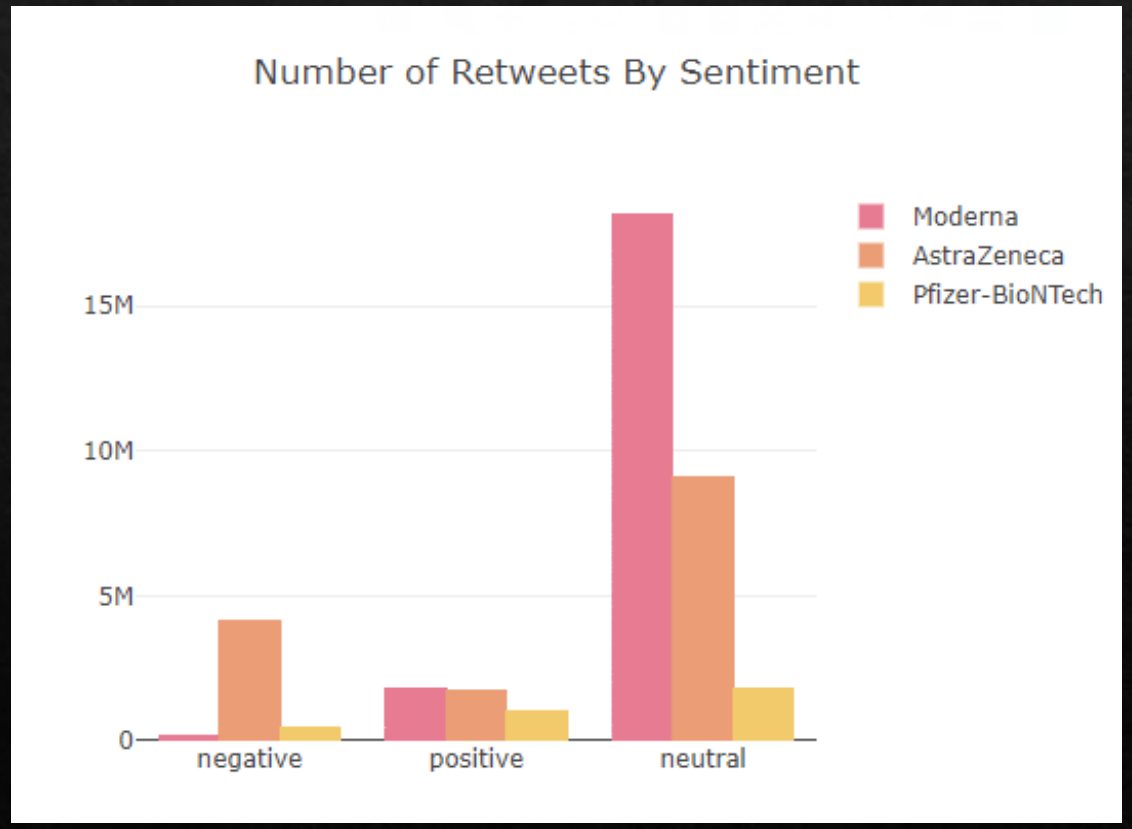
Pfizer-BioNTech

Tweet Polarity by Manufacturer



Retweets by Sentiment





Conclusions

There were more positive than negative tweets overall

Positive tweets were liked more than negative tweets regardless of company mention

Except for tweets which mentioned Astra Zeneca, positive tweets were retweeted more than negative tweets