S3715555 – Vikas Virani

Introduction

- My entity is “The New York Times (nytimes), An American newspaper based in New York with worldwide influence & readers”.
- The hypothesis being tested here is what are the current trending news & news topics that are being discussed and how people are reacting to those news (sentiments of people).
- As News is the prime source to get informed on current trends, discussions & international events, doing an analysis on topic modelling on the news provided by newspaper/news channel can help you identify what are the latest topics being discussed on international level.
- Also, People affected by those news are as important. So, by doing sentiment analysis on the people’s reaction/discussion on the news provided by New York Times can help identify the sentiments of people and how are they affected & what are the impacts of it on global level.

Data Collection

- As shown in the image below, I’ve fetched the first 800 tweets which included the search term “nytimes”, which will give us all the tweets by user which includes this term.
- Here, I’ve use REST API to fetch the tweets. Though it has a maximum of one week’s tweet limit, nytimes has a large followers & so I can get significant amount of tweets without hitting the limit.

```
In [1]: %reload_ext autoreload
        %autoreload 1
        %aimport twitterClient

In [2]: from tweepy import Cursor
        client = twitterClient.twitterClient()
        import json

In [3]: with open('tweets.json', 'w') as fJson:
            # retrieve the first 800 tweets
            for tweet in Cursor(client.search, q='nytimes', lang='en', tweet_mode='extended').items(800):
                # write out current tweet json to file
                json.dump(tweet._json, fJson)
                fJson.write('\n')
                #print(tweet._json['full_text'])
```

- I’ve retrieved 1200 tweets and stored it in a JSON file with each tweet on a new line amounting to total of 7.52 MB of JSON file. The format of retrieved tweet can be found here : https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json.html

- By using a print statement, it was observed that there are less amount of hashtags in the tweets fetched as the opinion on news was given by users in the tweet text itself without including hashtags. But total tweets being 1200 appears adequate to answer the questions.

Pre-processing and Data Cleaning

- For each Tweet in the JSON file, I've done following pre-processing:
    - There are inbuilt string punctuations in "string.punctuation", to denote which text is punctuation, "**stopwords**.words('english')" from **nltk** library to get the list of stopwords in English language & other Twitter specific words like "rt","via" etc.; all these are store In one list.

    - When tweet is tokenized with TweetTokenizer, each token is compared with this above list; if that word is from above list, it will be discarded from further processing.

    - A python file "TwitterProcessing" is created which is called from mail file to process tweet text every time. The file processes text by converting it into lowercase, removing leading & trailing **spaces**, **digits**, words starting with **http & @** to **remove links** & **user mentions**.

    - Additionally, it also normalizes word by lemmatization using "**WordNetLemmatizer()"** from nltk. Here lemmatization is used instead of stemming to normalize each words because unlike stemming it normalizes words to actual dictionary words, which will be useful in visualizing word clouds.

- Below image is a sample print of some tweets from the JSON file, as can be seen from the image, it contains RT, @ followed by username, links for related resources, starting with HTTP etc. Additionally, it contains many punctuations like ",", "?", ":" etc. & some words like "an", "the", "is" (stopwords) which are not useful in analysis; In order to remove all these things, I've done pre-processing on text to get clean text.

```
In [7]: with open('tweets.json', 'r') as f:
            for line in f:
                # each line is loaded according to json format, into tweet, which is actually a dictionary
                tweet = json.loads(line)

                tweetText = tweet.get('full_text', '')
                tweetDate = tweet.get('created_at')
                print(tweetText)
```

```
Miners Kill Indigenous Leader in Brazil During Invasion of Protected Land https://t.co/MCLXBAKc77
RT @nytopinion: "After I was elected prime minister of Pakistan last August, one of my foremost priorities was to work for last
ing and just…
@nytimes @fmanjoo LOL...oh dear. Sticks and stones. Offended, are you? If you take a stance, be prepared for conflict.
No, stop fuc*ing about. Address the big elephant: overpopulation.
RT @joshuawongcf: We understand that some critics of interventionism may be inclined to have sympathy for CN as a still-develop
ing country…
RT @nytimesworld: The Indian authorities are racing to build new prisons in Assam, some to house people as they are processed f
or deportati…
RT @Libertea2012: 51 People Died in Mass Shootings in August Alone in the U.S. https://t.co/UdhkSbSRLc
RT @FrankFigliuzzi1: Join us today at 7:15p ET: Michael Flynn's Lawyers Escalate Attacks on Prosecutors @MSNBC @JoyceWhiteVance
#flynn…
RT @AmbassadorJawad: While defending Kunduz against failed Taliban assault, we lost a brave defender, Col. Sayed Sarwar Hussain
i.  His sacr…
RT @nytimes: In Opinion
```

Analysis Approach

- The research was essentially based on two questions:

  1) Sentiments of People &
  2) Current Topics in trend, so I've performed two analysis to address these two questions; Sentiment Analysis & Topic Modelling.

- I've done **Vader sentiment analysis** to perform my analysis, using Vader SA instead of Opinion word Counting SA has its advantages;

  - Vader uses lexicon features like **punctuation**, **capitalisation**, **degree modifier** etc.; this help in identifying scores more accurately because some punctuations or capital words mean some specific emotion & some degree modifier hold more weight than other.

  - Apart from that, it scores words based on a scale of -4 to 4 instead of just 1 & -1 for positive & negative like Opinion word, which gives a weighted value of sentiments to get more accurate results. Hence it is more relevant in our case than Opinion word SA.

- I've used "**latent dirichlet allocation**" Topic Modelling in my analysis; because this algorithm assumes that each document is generated by picking a set of topics & then picking a set of words for each of these topics, it reverse engineers the process to identify topics.

  - As LDA is an iterative training, I've set **max_iter** to 10 to stop after 10 iterations. As we can give **number of topics** to be extracted & **number of words** to be consider to describe documents as parameters, it is a suitable approach to potentially identify different topics. I've used Number of **topics as 8** to distribute topics so that It can cover variations in news in different clusters

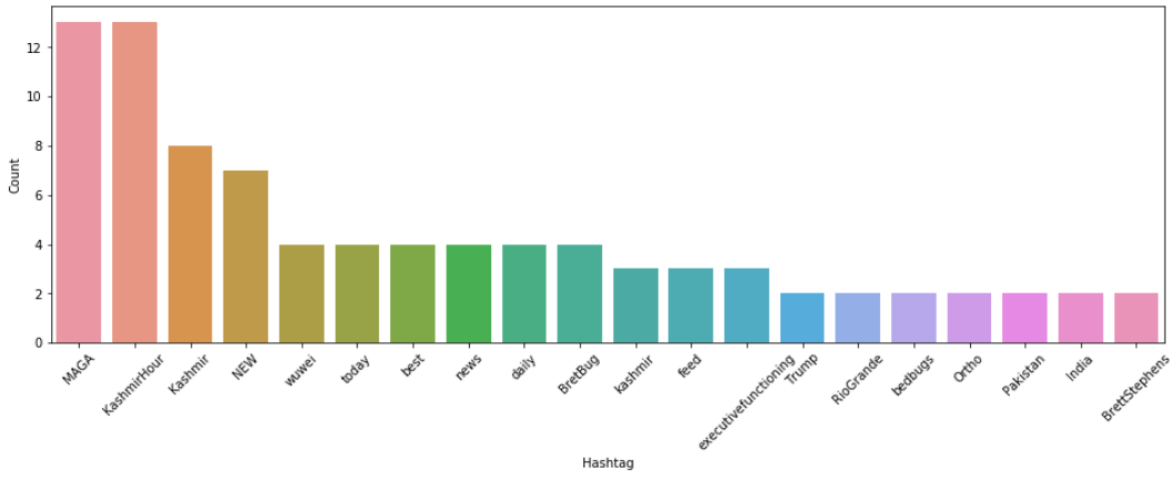
& number of **features/words as 2000** to cover all potential words to be considered

- Both these approaches are unsupervised learning approaches as we have no predefined set of labels available.

Analysis & Insights

- Though there were very less hashtags, I've extracted them & visualized in bar chart to show the trends of topic that are being discussed. Here is the bar chart;

  As can be seen from the graph, MAGA (Make America Great Again), Trump, India Pakistan conflicts for Kashmir, David Karpf calling Brett Stephens metaphorical bedbug are recent trending news.

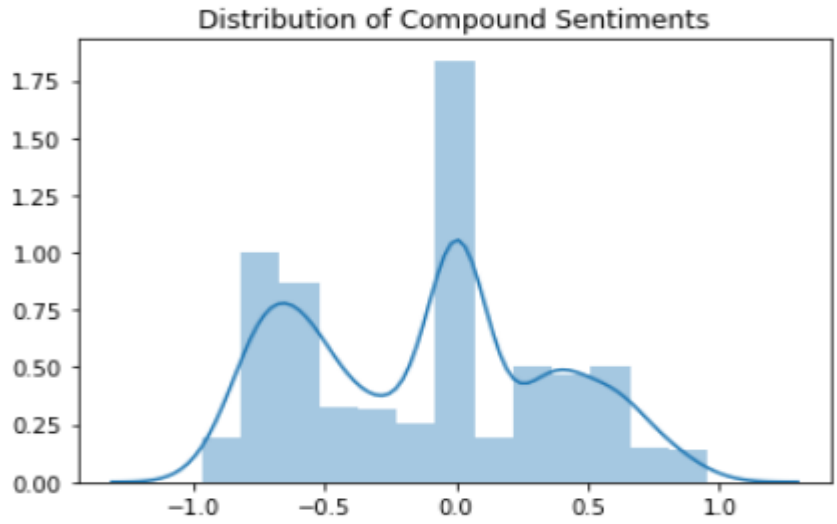


- From the Analysis of above two approaches;
  - People tend to have slightly negative sentiments on recent news by New York Times as can be seen from the statistics from below image; mean compound sentiment is a negative value which suggest negative sentiments. This can be the case maybe because of the recent trending news having bad influence.

Distribution of Sentiments

:

	Overall Sentiment	Positive	Negative
count	1200.000000	1200.000000	1200.000000
mean	-0.114376	0.121348	0.175098
std	0.472949	0.166669	0.193961
min	-0.967000	0.000000	0.000000
25%	-0.542300	0.000000	0.000000
50%	0.000000	0.000000	0.143000
75%	0.275500	0.219000	0.333000
max	0.955100	1.000000	1.000000

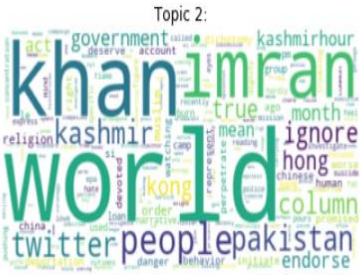


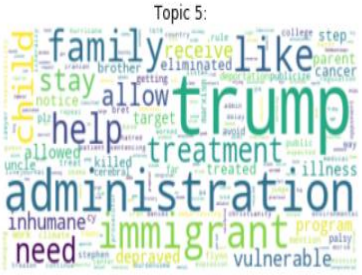
- From the LDA model, I've identified the words into 8 different topics; from which top 20 words from each topic are displayed as below;

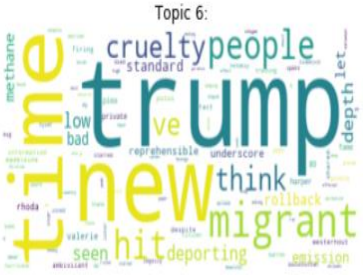
```
display_topics(ldaModel, tfFeatureNames, wordNumToDisplay)
```

```
Topic 0:
time opinion new york different prime minister day end woman like column article need year know op want thing ed
Topic 1:
world khan imran people pakistan twitter kashmir column ignore true hong kong government month endorse mean act kashmirhour religion order
Topic 2:
action trump family deferred isabel area medical personal maria bueso immigrant assistant madeleine westerhout year talk btw difference bay incl
Topic 3:
stephen bret indian bedbug book read kashmir world stop assault illiterate great want state functionally mattis jew quote nyt called
Topic 4:
trump administration immigrant like family help child treatment need stay allow vulnerable inhumane receive depraved program allowed illness target step
Topic 5:
trump new time migrant people hit cruelty ve think deporting seen low depth let methane bad emission standard rollback reprehensible
Topic 6:
people going kashmir human right expose violation ambassador oppression fascis gross new time york use life die le know editorial
Topic 7:
link bret followed trump kid come sentence death sign convinced company cheer horror maga beginnin major say regulation methane oil
```

- Below is the word clouds of each topic to know the frequency of words in each topic & better identify topic based on most occurring words.







- From these evidence, I was able to identify different topics that are currently being discussed i.e.:

    - Human rights abuse in Kashmir between India & Pakistan
    - David Karpf's controversial comment on calling Brett Stephens metaphorical bedbug
    - Trump's Policies for immigrants
    - Trump's rollback of Methane emission standards

All these topics were detected in word cloud, which matches to the actual current trends & topics or background knowledge that are being discussed in the world right now. As these topics are negatively influenced, people are affected negatively with these news emotionally & so the overall mean sentiment was negative which we derived from Vader SA.

- Since I've fetched the tweets from search option instead of user_timeline, the date stamp was same (the date when I retrieved the tweets) & so the time series plot of sentiment counts for each day was not possible. This plot is a good measure to identify people's sentiment over time for different topics.

Conclusion

- This Research was done to analyse the news topics currently trending & people's sentiment on those news by The New York Times.
- Vader Sentiment Analysis approach was used to identify sentiments of people & latent dirichlet allocation (LDA) was used to identify topics/trends being discussed.
- From the analysis of the results from above two approaches, it is observed that people have negative sentiments for recent news because recent news topics are negatively affecting people's emotions one way or another & topics/trends Identified by LDA corresponds to actual current topics/background knowledge.