Topic Modeling and Sentimental Analysis of Tweets on Covid19 to Find the Weightage of the Popular Hashtag

J. Jeyasudha[a] and Dr.G. Usha[b]

[a] Research Scholar, Department of Computer Science and Engineering, SRM Institiute of Science and Technology, Tamilnadu, India. jjeyasudha@gmail.com

[b]Associate Professor, Department of Software Engineering, SRM Institiute of Science and Technology, Tamilnadu, India.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: The Hashtags plays a vital role in the social media and it is easily highlighted by each and every people when they tag it for their own views. Marketing and advertisement is booming so that to make their products work through the views of the normal or common people. Sometimes they use the false content for their publicity and misleading the people. In this paper, the covid19 tweets are taken for finding out the popular hashtags using the correlation techniques like pearson, spearman and kendall rank correlation. The Covid19 hashtag is more popular with the correlation coefficient and sentimental analysis of the tweet than coronovirus tag. To justify the popularity, the weightage's of the hashtag is found out by applying the topic modeling. In that the coronovirus tag is having more weightage than Covid19 tag.

Keywords: Topic Modeling, Sentimental Analysis, Correlation Coefficient.

1. Introduction

Hashtags, an undeniable happening occurs in our everyday routine in social media account(s) either as a tweet from a group or as a campaign or an event. They are always there, as they pop up when we do watch an advertisement, movie, drama etcetera. It is obvious and a challenge to users to obtain the correct information and to identify the right user for a right post. Literature shows that, there are many tools developed to track social media which in turn gives an opportunity for users to make decisions over a hashtag post.

Even, there are tools developed to track posts in social media networks[1,2], at times, users may get lost in recognizing correct information and may miss the strategic point of view on a hastag post. This scenario puts every user to think exactly what information to be retrieved from the report, such as number of tweets along with retweet's, links and original tweets, who posted what tweet and their relevancy and ranking of the users. In the present work, it is proposed to explain basic social media manipulations that are necessary in any of the social media hashtag analysis.

1.1. Original Tweets and Retweets

Our world runs on online marketing, very important concept to survive and succeed in online marketing. Twitter has a unique feature called "retweet". The very basic point to observe on a Twitter post for a user is to understand the difference between Tweets and Retweets. So that, retweets can be separated from the original tweets when in need to run a good Twitter analytic report. This provides an ambient opportunity to know the where it differ's through an systematic point of view. On the other hand, understanding twitter typification of tweets results positively in a hashtag counter. The general categorization of tweets is: Tweets and Retweets.

Original Tweets

Original tweets are the one provides actual content. This indicates that, the tweets are written from the original thoughts by users. These original tweets provide value to the content in a hashtag or makes the tweets are analysed. Further, these original tweets can be classified as Text Tweet, Replies and Links and Pictures.

Retweets

Retweets are duplication of original tweets that have already been twitted. More often, they will not add any new significant information to the hashtag whether it is a picture or to a link. Hence, they always will be a retweet.

What is Better-Tweets or Retweets

It is a tricky question and there is only one answer to it. Both tweets make impact on a hashtag post which is shared and discussed. Sometimes, retweets possesses more interest than original tweets. Hence, the less retweets the better original content that the report has but not in the case of online marketing. As an example, a high influence account user retweets a meaningless post may have a huge impact on the report. Either way, it is important to distinguish between tweets and retweets to define the happenings of marketing strategies. Twitter considers both tweets and retweets as "Tweets". Hence, the total number of tweets from a hashtag report is really a mixed one. The advanced searching helps user in identifying mix level and its influence on the hashtag report. As an example, when typing #You #Pic, displays all the tweets that are pictures and that contain the hashtag. Besides the typology of tweets, the crucial part is to understand the marketing strategy is engaged with the user or it is only collecting data (RTs) to the tweets owned by the main account or brand.

Twitter Impressions vs Twitter Reach

Twitter impact is based on how many times the posted hashtag has been seen by Twitter Users whereas Twitter Reach refer to how many users have seen the posted hashtag.

How many Twitter Impressions is good?

Tweet Impressions: The followers have 20% impression will be good. But 20% usually changes. It is noted that 20% of the followers saw the tweet and not to be followers alone. Hashtags Impressions: No fixed number. In Figure.[1.1] shows the frequency of tweets depends on day to day impressions of the followers.

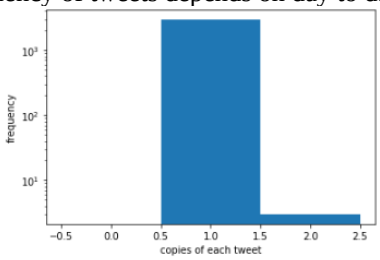

Figure 1.1. The frequency of tweets about covid19

2. Keyword Correlations In Text

The retweeted tweets have been removed, referenced tweets and the different section hashtags of their own. The important part to be realized is exceptionally retweeted person, referenced person and the famous hashtags shown in Figure.[1.2] that is been circulated. In the accompanying segment the play out of examination on the hashtags as it were. The surrender of tweets over to the person in return and rehash a comparative examination on the referenced and retweeted segments. The foremost first part is to choose the segment of hashtags from the panda's dataframe, and the lines that are take in which hashtag is really available.

	hashtags
1	[#COVID19, #COVID19]
8	[#BeBuried]
12	[#SenatorForSale]
22	[#COVID19]
25	[#Coronavirus]

	hashtag	counts
0	#COVID19	549
1	#coronavirus	232
2	#Covid19	72
3	#Coronavirus	69
4	#COVID	20

Figure 1.2. The popular hashtags and usage counts

2.1. From Corona dataset Tweets Text to the Vector

Presently a discovery of the hashtags that are related with one another is most needed. To perform a transformation of the content into numeric structure is to be done. The more conceivable of changing from a rundown of tweet's hashtags to a vector speaking to the hashtags showed up in which lines. In the event that accessible hashtags were the set [#geography, #cuties, #sunny, #date], at that point the tweet '#sunny #cuties' would be [0,1,1,0] in vector structure. The technique is applied to the hashtags segment of df. Before that it is confine to tweet hashtags that show up finite occasions to be connected with different tweet hashtags. The tweet hashtags are not related with tweet shown up once, and there is no need of tweet hashtags that seem a small number of times because it could prompt deceptive relationships. The discovery of tweet hashtags meet a base edge. The next part is that channels the tweet hashtags to just the well known tweet hashtags. There is drop of the lines when no famous hashtags is available dropping of the the popular_hashtags column from the dataframe. The correlation of the dataframe columns and the correlation between the different hashtags appearing in the same tweets in the Figure. [2.1].

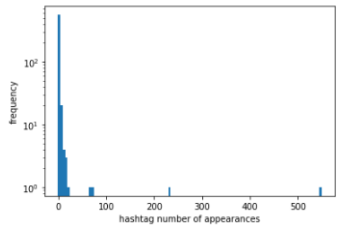

Figure 2.1. The frequency of the popular hashtag's

Correlation is a bivariate research that gauges the characteristics of dating among factors. In insights, the estimation of the relationship coefficient adjustments among +1 and - 1. At the factor whilst the estimation of the connection coefficient lies around ± 1, at that factor it is meant to be a really perfect degree of connection among the two factors. As the connection coefficient esteem goes towards 0, the connection between the two factors will be more fragile. The three types of correlations: Pearson correlation, Kendall rank correlation and Spearman correlation. Pearson correlation is measure of the degree of the association involving linear related variables and mostly used for research [7,8]. And the pearson coefficient between the covid19 hashtag and coronavirus falls under the category of 0.0 to 0.4 as in Figure. [2.2].

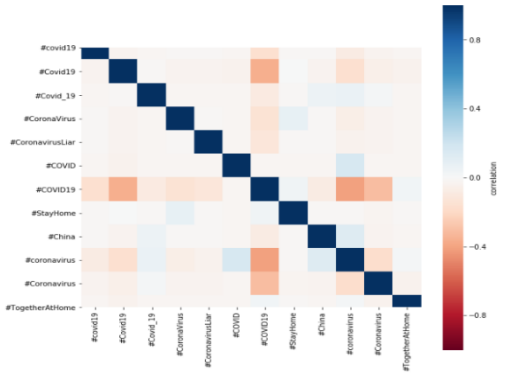

Figure 2.2. Pearson correlation of the covid19 tweets

Spearman rank correlation could be a non-parametric evaluation utilized to quantify the amount of relationship between the factors. It had been created by Spearman, on these lines it's called the Spearman rank relationship. Spearman rank affiliation take a look at does not settle for any presumptions regarding the dissemination of the knowledge and is that the appropriate relationship examination once the factors square measure calculable on a scale that's at any rate ordinal [9,10]. And the spearman coefficient between the covid19 hashtag and coronavirus falls under the category of 0.0 to 0.4 as in fig. [2.3]. Additionally usually known as "Kendall's tau coefficient". Kendall's Tau coefficient and Spearman's rank relationship coefficient survey factual affiliations dependent of the information on the positions. Kendall rank relationship is an option in contrast to Pearson's connection (parametric), the information of working with has bombed at least one presumptions of the test. It can be better alternative to Spearman relationship (non-parametric) if the example size is little and has numerous tied positions.

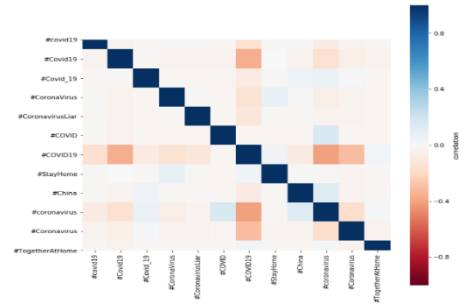

Figure 2.3. Spearman correlation of the covid19 tweets

Kendall rank connection is utilized to check the likenesses of information request when it is positioned by amounts. Different sorts of connection coefficients utilize the perceptions as the premise of the relationship, Kendall's connection coefficient utilizes group of perceptions and manipulates quality affiliated and dependent on the pattern on concordance and dissonance between the groups. And the coefficient between the covid19 hashtag and coronavirus falls under the category of 0.0 to 0.4 as in fig. [2.4].

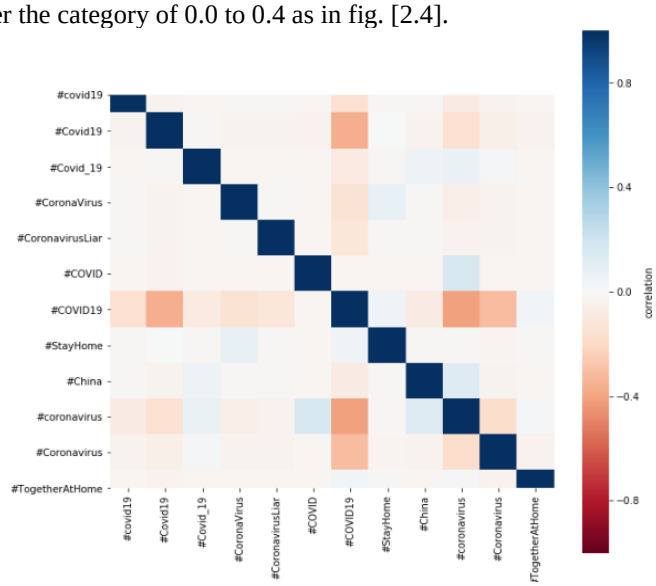

Figure 2.4. Kendall's tau coefficient correlation of the covid19 tweets

3. Cleaning Unstructured Text Data

Displaying the calculation of the subject is to based upon the tweets. On the off chance tweets that are confused, with non-standard English, capitalization, joins, tweet hashtags, @users and accentuation and emojis all over. In the event that it will be ready to apply subject demonstrating it have to evacuate a large portion of this and back rub our information into a progressively standard structure before at last transforming it into vectors. In this area they will give a few capacities to cleaning the tweets just as the explanations behind each progression in clearout. The utilization of the nltk python bundle and change to appropriately and proficiently clearout content information will be other full instructional exercise so the trust is sufficient to remove. The imports of the tweets is the first foremost task.

In the cell beneath we have given you a few capacities to expel web-joins from the contents. There is no explicit web connections for the significant data, in spite of the fact that in the event that you needed to provide supplant all link joins with a token [9,10], The data that is saved and a web interface to safeguarding the connection itself. For the situation in any case, we will evacuate joins. The evacuate of retweets and makes reference to the method. In fact that it is impossible that they will assist us with shaping significant points. The master function is for the some more cleaning of the data using two functions.

The area on visual cues depicts what the tweet cleaned ace capacity is making a progression. In the event that you need you can skip perusing this segment and simply utilize the capacity until further notice. You will probably see some unusual contents of the area specified, at last produce them in return to second last visual cue about stemming and applied these means all together.

- Remove the clients from connections and in the tweets however tweet hashtags as we accept that in any case mention to us what individuals are discussing in a progressively broad manner.
- The entire tweet lowercase is used for calculation that is equivalent.
- The part is to expel accentuation letters, available in the my_punctuation string, to additionally clean up the content. Discover tokens in the accentuation toward the last or in the center.
- Evacuate twofold separating by the accentuation expulsion with expel numbers.

4. Sentimental Analysis Of The Covid Datset To Find Popular Tag

The number of tweets about Covid-19 and Coronavirus are 628M tweets up until now.Perhaps the saddest thing we have ever needed to do is to dissect the advancement of the new Coronavirus Covid 19 on Twitter. We will refresh the information consistently and we will give as much data as possible. On the off chance that you are a columnist you can utilize this unreservedly by referencing Tweet Binder as a source. On the off chance that you need the crude information (the tweets), if you don't mind get in touch with us since that can't be offered for nothing. Coronavirus on Twitter is likewise brimming with counterfeit news, if it's not too much trouble remain sharp.Since many individuals are tweeting about Coronavirus, we have run a hunt with all the tweets that got 1000 RTs at any rate. We consider those "significant", that doesn't mean clearly that they are the most significant tweets of the Coronavirus issue. There are in excess of 40,000 tweets about Covid19 who got in any event 1,000 RTs in a tweet.

Topic modelling on the covid19 dataset is to find persons tweeting about covid19.An example dataset the content information is cleaned and investigate well-known hashtags are being utilized, the persons tweeted and retweeted, lastly we will utilize two solo AI calculations, explicitly idle dirichlet portion (LDA) and non-negative network factorisation (NMF),[9,10] to investigate the subjects of the tweets in completely.

Twitter is an incredible wellspring of information in researcher point of view, more than 8,000 tweets sent for each second. The contents that a huge number of clients send can be used and broke down to attempt to explore group feeling on specific issues. A fundamental is about searching for watchwords and expressions. The part next is to discover persons tweeting and no more, the most retweeted, and the most widely recognized hashtags.

- Persons tweeting
- persons tweeted at/referenced
- The hashtags are being utilized

	hashtags
1	[#COVID19, #COVID19]
8	[#BeBuried]
12	[#SenatorForSale]
22	[#COVID19]
25	[#Coronavirus]

	hashtag	counts
0	#COVID19	549
1	#coronavirus	232
2	#Covid19	72
3	#Coronavirus	69
4	#COVID	20

Figure 3.1. The popular hashtags and usage counts

The person retweeted are removed/referenced and different sections hashtags. Fig. [3.1] confirms that there are more tag Covid19 and coronovirus are more popular. The positive and negative values are computed for

each and every corona tweets and been maintained in the dataframe [3,4]. And the sentimental analysis is done for the tweets before finding the popular hashtag's, some of the tweets are given below with the positive and negative value and probability as in fig. [3.2].

```
Predictions:

Tweet: The #Coronavirus death count at 46,000 is completely untrustworthy. In a single day, New York added almost 4,000 people—
nearly 10% of the total—who never tested positive. With nationwide shenanigans, encouraged by the CDC, one can only guess at th
e real death count. Shameful!
Predicted sentiment:  Positive
Probability:  0.94

Tweet: Two cats in New York have been infected with the novel coronavirus, making them the first pets in the US known to be inf
ected, federal officials say. Both had mild respiratory symptoms and are expected to make a full recovery. https://t.co/nCNFPEt
GrC
Predicted sentiment:  Positive
Probability:  0.96

Tweet: The @WHO is now saying that due to a potential second resurgence of #COVID19 in the winter, we should all get flu shots
this year.
Predicted sentiment:  Negative
Probability:  0.69

Tweet: Does the flu shot somehow protect against COVID now?
Predicted sentiment:  Positive
Probability:  0.57

Tweet: BREAKING: UK #coronavirus hospital deaths rise by 763 in past 24hrs, bringing total to 18,100.
Predicted sentiment:  Positive
Probability:  0.53
```

Figure 3.2. The positive and negative statements of corona tweets with probability

5. Applying Topic Modeling

The software are used to clean tweets and to transform the content tweets into vectors and a model is constructed. To transform the content into a matrix, it encodes each line in the grid and the tweets used in individual tweet. The words max_df=0.9 implies that the tweets and words >90% of tweets. The words max_df=0.25 implies that the tweets and words >25% of tweets , 25 tweets will be disposed of. We dispose of high showing up words since they are too normal to be in any way significant in points. We dispose of low showing up words since we won't have a sufficient sign and they will simply acquaint clamor with our model. We typically transform content into a meager network, save the space, the database utilize an ordinary grid. Look at the state of tf (term recurrence - the recurrence of each letters in each tweet). The state of tf reveals to us the tweets counts and the words to be sifting procedure. The tf_feature_names [13] to perceive what tokens endured separating. The tf network is actually similar to the hashtag_vector_df dataframe.A word is available in each line of the tweet

The subjects are picked for the model items. The irregular state is characterized with the goal is not producing good model. The model is our LDA calculation model article. We expect that on the off chance that you are here, at that point you ought to be OK with Python's item direction. The point to know is that the model item is needed. The parameters like the quantity of subjects holds when we made it; it additionally holds strategies the fitting strategy; when we fit it, it will hold fitted parameters which reveal to us how significant various words are in various points [5,6].

The request for the words in our framework tf_feature_names and the quantity of words we might want to appear. Utilize this capacity, which restores a dataframe, to show you the subjects we made. Recollect that every subject is a rundown of words/tokens and loads Presently [11,12] we have a few themes, which are simply groups of words, we can attempt to make sense of what they truly mean as mentioned in the Figure.[5.1].

	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights	Topic 3 words	Topic 3 weights	Topic 4 words	Topic 4 weights	Topic 5 words	Topic 5 weights
0	#covid	282.3	coronaviru	96.9	death	171.2	coronaviru	106.7	'	325.7	coronaviru	232.6
1	case	140.6	die	84.7	week	135.0	death	95.3	coronaviru	138.3	trump	121.2
2	#	92.1	hospit	71.6	coronaviru	106.5	home	88.2	it	76.1	china	95.8
3	april	85.6	covid	51.7	ago	94.1	peopl	72.1	like	57.3	hous	70.1
4	test	85.6	protect	50.1	case	83.4	protest	71.7	china	54.2	white	58.1
5	ke	62.1	need	47.6	million	68.5	stay	55.1	say	54.0	presid	51.0
6	#coronavirus	54.5	#covid	47.3	report	66.1	governor	47.9	time	51.5	brief	49.1
7	today	52.8	help	43.5	us	62.4	state	46.4	live	51.2	lab	49.0
8	pm	49.0	patient	40.5	u	56.1	day	46.1	uk	47.5	democrat	49.0
9	number	47.5	work	38.6	relief	30.0	break	46.0	respons	46.6	govern	48.3

Figure.5.1 Topic modeling of Covid Dataset

6. Conclusion

The proposed work focuses on the covid19 dataset and the popular hastags have been forund out using sentimental analysis amoung the 8,000 tweets, 6139 tweets are off positive and remaining tweets are the negative comments. And the topic modeling has been done on the dataset and forund the wethtage that has been given for each and every word. In that we found out that coronavirus has the more weightage in all sort of sets it is of more than 100.

References

1. K. Anand, G. Bianconi, "Entropy measures for networks: Toward an information theory of complex topologies", Physical Review E, Vol. 80, No. 4, pp. 1-4, 2009.
2. S. Cao, M. Dehmer, "Degree-based entropies of networks revisited", Applied Mathematics and Computation, Vol. 261, pp. 141-147, 2015.
3. T. Nie, Z. Guo, K. Zhao, Z.M. Lu, "Using mapping entropy to identify node centrality in complex networks", Physica A: Statistical Mechanics and its Applications, Vol. 453, pp. 290-297, 2016.
4. L. Fei, Y. Deng, "A new method to identify influential nodes based on relative entropy", Chaos, Solitons & Fractals, Vol. 104, pp. 257-267, 2017.
5. S. Peng, A. Yang, L. Cao, S. Yu, D. Xie, "Social influence modeling using information theory in mobile social networks", Information Sciences, Vol. 379, pp. 146-159, 2014.
6. H.W. Shen, "Community structure of complex networks", Springer Science & Business Media, 2013.
7. Lancichinetti, S. Fortunato, F. Radicchi, "Benchmark graphs for testing community detection algorithms", Physical review E, Vol. 78, No. 4, pp. 1-6, 2008.
8. M.E. Newman, "Assortative mixing in networks", Physical review letters, Vol. 89, No. 20, pp. 1-5, 2002.
9. N. Litvak, R. Van Der Hofstad, "Uncovering disassortativity in large scale-free networks", Physical Review E, Vol. 87, No. 2, pp. 1-11, 2013.
10. B. Min, F. Liljeros, H.A. Makse, "Finding influential spreaders from human activity beyond network location", PloS one, Vol. 10, No. 8, pp. 1-13, 2015.
11. L. Lü, L. Pan, T. Zhou, Y.C. Zhang, H.E. Stanley, "Toward link predictability of complex networks", Proceedings of the National Academy of Sciences, Vol. 112, No. 8, pp. 2325-2330, 2015.
12. F. Morone, H.A. Makse, "Influence maximization in complex networks through optimal percolation", Nature, Vol. 524, pp.65-68, 2015.
13. J. Jeyasudha, G. Usha, "Community Spam Detection Methodologies for Recommending Nodes", International Journal of Recent Technology and Engineering (IJRTE), Vol.8, No.2S4, pp. 131-142, 2019.