ABSTRACT

Nowadays, people from all over the world use social media platforms to share information. Twitter, for example is a platform in which users send, read posts known as ‘tweets’ and interact with different communities. Users share their daily lives, post their opinions on everything such as brands and places. Companies can benefit from this massive platform by collecting data related to opinions on them. The aim of this work is to present an approach that can perform sentiment analysis of real data collected from Twitter. Data in Twitter is highly unstructured which makes it’s analysis difficult. The process of performing sentiment analysis as follows: Tweet extracted from Twitter API, then cleaning and discovery of data performed. After that the data was fed into the approach for the purpose of analysis. Each tweet extracted was classified based on its sentiment whether it is a positive, negative or neutral. Tweets were collected on following subjects:

- LockDownIndia
- Covid19India
- CoronaIndia
- quarantine
- lockdown
- lockdown1
- lockdown2
- lockdown3
- IndiaFightsCorona
- india_against_covid19

The approach used was Rule-based sentiment analysis using Natural Language Processing (NLP). The result from this approach was shown using data visualization methods.

1. Introduction

The online social media such as Twitter, Facebook, and Instagram allow users to communicate with the whole world. Write their own opinions about products or share their moments, even influence politics and companies. Twitter for example, almost every huge company have an account on Twitter to know about their customers feedback about their services or products. Sentiment analysis, known as opinion mining, for classifying specific words into positive or negative[1]. In this work, sentiment analysis has been used to classify specific English tweets about three lockdowns in India due to covid19. The task is to determine whether specific tweet is positive, negative or neutral.

In order to analyse it better, the concept of topic modelling has been used that shows the topics that dominated the tweets. **Topic Modelling** is the task of using unsupervised learning to extract the main topics (represented as a set of words) that occur in a collection of documents.

Latent Dirichlet Allocation (LDA) is used to classify text in a document to a particular topic. It builds a topic per document model and words per topic model, modelled as Dirichlet distributions.

- Each document is modelled as a multinomial distribution of topics and each topic is modelled as a multinomial distribution of words.
- LDA assumes that every chunk of text we feed into it will contain words that are somehow related.
- It also assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution[2].

In its simplest form, Sentiment analysis is modelled as a set of two classification problems:

Subjectivity – classifying a sentence as subjective or objective, known as subjectivity classification.

Polarity – classifying a sentence as expressing a positive, negative or neutral opinion, known as polarity classification.

Rule-based Systems

Usually, rule-based systems run on a set of rules that identify subjectivity, polarity etc. We can summarize a basic rule-based system in following steps:

- Define two lists of polarized words negative words and positive words. Some examples of negative words are awful, bad, worst, etc. and some positive words are beautiful, good, best, etc.
- Given a text:
  Count the number of positive words in the text.
  Count the number of negative words in the text.
  If the number of positive words is greater than the number of negative words then return a positive sentiment, conversely, return a negative sentiment. Otherwise, return neutral[3].

2. Methodology and Results

This work focusses on mining tweets written in English. The number of tweets extracted were 124384. The roadmap followed for analysing sentiments of tweets is as follows:

Step 1: Basic reading of data

This includes importing the data.

	id	conversation_id	created_at	date	time	timezone	user_id	username	name	place	
0	1.262790e+18	1.262790e+18	1.589910e+12	19-05-2020	16:51:14	UTC	1.250080e+18	dramaflick	The Drama Flick	NaN	https://www.youtube.com/wat CRbO7ExO
1	1.262790e+18	1.262790e+18	1.589910e+12	19-05-2020	16:50:44	UTC	8.078430e+17	knowpuneet	TravelTrainee	NaN	Lockdown 4.0 ka naam hi loc hai\nHa
2	1.262790e+18	1.262790e+18	1.589910e+12	19-05-2020	16:48:29	UTC	1.085430e+18	narasinhpurohit	Narasinh Purohit	NaN	CORONA VIRUS THREAT-\nH OVERCOME STRE
3	1.262790e+18	1.262790e+18	1.589910e+12	19-05-2020	16:47:36	UTC	1.104210e+18	ka_trolls	Humans Of Hindutva	NaN	Co please\n\n#lockdownindia\n@Bhuva
4	1.262790e+18	1.262790e+18	1.589910e+12	19-05-2020	16:47:29	UTC	3.921802e+08	rajendrabohora	rajendrabohora	NaN	In fight with #COVID19, You are t

5 rows × 34 columns

Fig 1: Dataframe output

Step 2: Neglecting redundant columns and considering relevant columns for analysis

Various columns contain missing values and many columns are not so relevant for analysing sentiments. Hence, those columns have been neglected and our actual data will contain only two columns that is ‘tweet’ and ‘hashtags’ as shown in fig 2.

	tweet	hashtags
0	https://www.youtube.com/watch?v=-CRbO7ExO1k	['#lockdownindia', '#lockdown', '#indiafightsc...
1	Lockdown 4.0 ka naam hi lockdown hai\nHai sab ...	['#lockdownindia', '#locldown4']
2	CORONA VIRUS THREAT-\nHOW TO OVERCOME STRESS A...	['#covid_19', '#covid_19sa', '#covid_19india',...
3	Could you please\n\n#lockdownindia\n@Bhuvan_Ba...	['#lockdownindia', '#roastchallenge', '#journa...
4	In fight with #COVID19, You are the best Docto...	['#covid19', '#coronavirus', '#patiencechallen...
...	...	...
124379	I pledge to follow the appeal given by Hon'ble...	['#staysafestayhome', '#janta_curfew', '#janta...
124380	Four new cases of Coronavirus detected in Luck...	['#coronaindia', '#coronavirusoutbreakindia', ...
124381	Do you sometimes feel a tingling #sensation or...	['#sensation', '#hands', '#thevoiceofwoman', '...
124382	Some Time we have to Stay Back ... Just to Sav...	['#stayback', '#gobackcorona', '#coronafighter...
124383	Sir, National Medical Emergency should declare...	['#coronaindia', '#coronavirusupdate', '#wewil...

124384 rows × 2 columns

Fig 2: Final Dataframe

Step 3: Getting basic information of the data.

```
---Print the basic info of the data-----
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 124384 entries, 0 to 124383
Data columns (total 2 columns):
tweet       124384 non-null object
hashtags    124384 non-null object
dtypes: object(2)
memory usage: 1.9+ MB
None
(124384, 2)
tweet       object
hashtags    object
dtype: object
```

Fig 3: Dataframe basic info

Step 4: Pre-processing the data

As the text is highly dimensioned unstructured data, it has to be cleaned and prepared first before analyzing it. Pre-processing the data involves many tasks, depending on type of analysis. In this case, text has been extracted from tweets and converted to data frame, removed URLs from text, removed stop words like (the, a, to...), usernames and accounts (mentions), removed punctuations and converting encoding (Emojis), removing hashtags and retweets. Fig. 4 shows the cleaned tweet and hashtags columns respectively.

	tweet	hashtags
0	... THE DRAMA FLICK presents LOCKDOWN 4 that ...	['lockdownindia', 'lockdown', 'indiafightscovi...
1	Lockdown 4.0 ka.naam hi lockdown hai Hai sab u...	['lockdownindia', 'locldown4']
2	CORONA VIRUS THREAT- HOW TO OVERCOME STRESS AN...	['covid_19', 'covid_19sa', 'covid_19india', 'l...
3	Could you please lockdownindia _Bam RoastC...	['lockdownindia', 'roastchallenge', 'journalis...
4	In fight with COVID19, You are the best Doctor...	['covid19', 'coronavirus', 'patiencechallenge'...
...	...	...
124379	I pledge to follow the appeal given by Hon'ble...	['staysafestayhome', 'janta_curfew', 'jantacur...
124380	Four new cases of Coronavirus detected in Luck...	['coronaindia', 'coronavirusoutbreakindia', 'c...
124381	Do you sometimes feel a tingling sensation or ...	['sensation', 'hands', 'thevoiceofwoman', 'cor...
124382	Some Time we have to Stay Back ... Just to Sav...	['stayback', 'gobackcorona', 'coronafighters',...
124383	Sir, National Medical Emergency should declare...	['coronaindia', 'coronavirusupdate', 'wewillfi...

124384 rows × 2 columns

Fig 4: Cleaned tweets

Step 5: Find Subjectivity and Polarity scores

	tweets	Subjectivity	Polarity
0	... THE DRAMA FLICK presents LOCKDOWN 4 that ...	0.540000	0.160000
1	Lockdown 4.0 ka.naam hi lockdown hai Hai sab u...	0.288889	-0.155556
2	CORONA VIRUS THREAT- HOW TO OVERCOME STRESS AN...	0.500000	0.500000
3	Could you please lockdownindia _Bam RoastC...	0.000000	0.000000
4	In fight with COVID19, You are the best Doctor...	0.650000	0.333333
...	...	...	...
124379	I pledge to follow the appeal given by Hon'ble...	0.000000	0.000000
124380	Four new cases of Coronavirus detected in Luck...	0.454545	0.136364
124381	Do you sometimes feel a tingling sensation or ...	0.166667	-0.166667
124382	Some Time we have to Stay Back ... Just to Sav...	0.000000	0.000000
124383	Sir, National Medical Emergency should declare...	0.425000	0.000000

Fig 5: Subjectivity and Polarity scores

Step 6: Visualizing wordcloud of tweets

Generating word cloud from text gives more sense about the most frequently words used in tweets about a specific topic as shown in fig. 6. It is done by using WordCloud module of nltk package.

Fig 6: Word cloud from text

Step 7: Compute negative (-1), neutral (0) and positive (+1) analysis

In this phase, after preparing tweet (removing unnecessary symbols), each tweet was labelled as 1, -1, 0. (That's it: positive, negative, or neutral). Each word within each tweet is compared to positive and negative documents in order to find matching words, and classify tweets whether it has more positive or negative words.

	tweets	Subjectivity	Polarity	Analysis
0	... THE DRAMA FLICK presents LOCKDOWN 4 that ...	0.540000	0.160000	Positive
1	Lockdown 4.0 ka.naam hi lockdown hai Hai sab u...	0.288889	-0.155556	Negative
2	CORONA VIRUS THREAT- HOW TO OVERCOME STRESS AN...	0.500000	0.500000	Positive
3	Could you please lockdownindia _Bam RoastC...	0.000000	0.000000	Neutral
4	In fight with COVID19, You are the best Doctor...	0.650000	0.333333	Positive
...	...	...	...	...
124379	I pledge to follow the appeal given by Hon'ble...	0.000000	0.000000	Neutral
124380	Four new cases of Coronavirus detected in Luck...	0.454545	0.136364	Positive
124381	Do you sometimes feel a tingling sensation or ...	0.166667	-0.166667	Negative
124382	Some Time we have to Stay Back ... Just to Sav...	0.000000	0.000000	Neutral
124383	Sir, National Medical Emergency should declare...	0.425000	0.000000	Neutral

Fig 7: Analysis table

Step 8: Find bag of words

This involves the following: Pre-processing the raw data.

- **Tokenization**: Split the text into sentences and the sentences into words. Lowercase the words and remove punctuation.
- Words that have fewer than 3 characters are removed.
- All **stopwords** are removed.
- Words are **lemmatized** — words in third person are changed to first person and verbs in past and future tenses are changed into present.

Prior to topic modelling, we convert the tokenized and lemmatized text to a bag of words — which you can think of as a dictionary where the key is the word and value is the number of times that word occurs in the entire corpus[2]. In this case, we have:

```
Dictionary(187147 unique tokens: ['4', 'button', 'come', 'drama', 'fighting']...)
```

Frequency of each word can also be computed using FreqDist of nltk as shown in fig. 8.

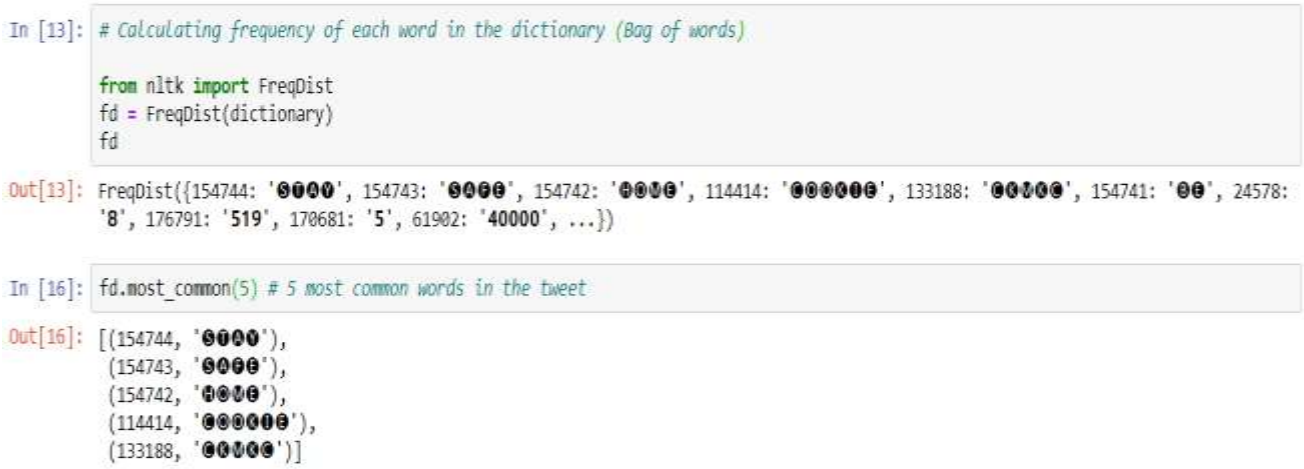

Fig 8: Bag of words and their frequency in the tweets

Step 9: Find main topics in the tweets

After computing the term frequencies, we can visualize main topics in the tweets for a better analysis.

	Topic # 0	Topic # 1	Topic # 2	Topic # 3	Topic # 4	Topic # 5	Topic # 6	
0	lockdownindia	recovered	coronaindia	narendramodi	coronavirusupdate	protect	coronaindia	india
1	people	people	total	curfewinindia	jantacurfew	23	stay	coron
2	time	india	corona	kerala	curfew	lockdownindia	home	coronavi
3	u	govt	covid19	tamilnadu	21dayslockdown	quarantine	safe	
4	please	state	case	chennai	jantacurfewchallenge	lockdown	social	
5	like	government	covid19india	pmmodi	indialockdown	236	coronacrisis	
6	take	r	coronavirus	22	fund	indiavscorona	doctor	stayh
7	one	lockdownindia	update	hantavirus	isolation	covid19	distancing	
8	corona	2	india	bengaluru	gocorona	21dayschallenge	mask	
9	need	1	coronavirusoutbreakindia	narendermodi	whatsapp	video	please	corc
10	lockdown	indiafightcorona	coronaupdatesinindia	gocoronacoronago	besafe	fightagainstcorona	hand	
11	help	country	coronaviruspandemic	pmofindia	lockdownquery	retweet	staysafestayhome	lo
12	sir	case	coronavirusindia	in	kanika	quarantinelife	socialdistancing	
13	let	day	death	midnight	pic	...	coronainmaharashtra	
14	know	italy	latest	ckmkb	22nd	watch	coronavillains	st
15	day	3	last	announces	forward	workfromhome	fight	stayatho
16	coronaindiacoronaindia	testing	active	vegetable	example	indiafightscoronavirus	covidiot	
17	go	coronaindiacoronaindia	confirmed	jai	gocoronago	lockdown4	mumbai	
18	life	china	coronavirusoutbreak	elder	indiafightscornona	join	dear	cor
19	good	due	see	pmo	yesterday	war	police	soc

Fig 9: Main topics in each tweet

Step 10: Visualizing positive, negative and neutral sentiments distribution through various curves.

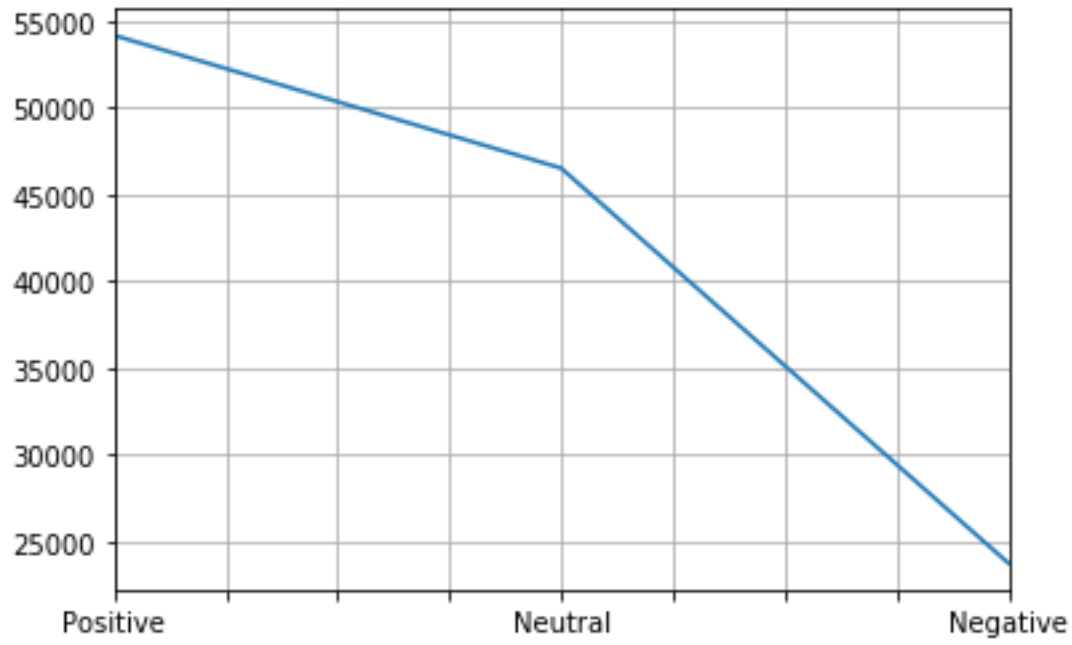

Fig 10: Analysis curve

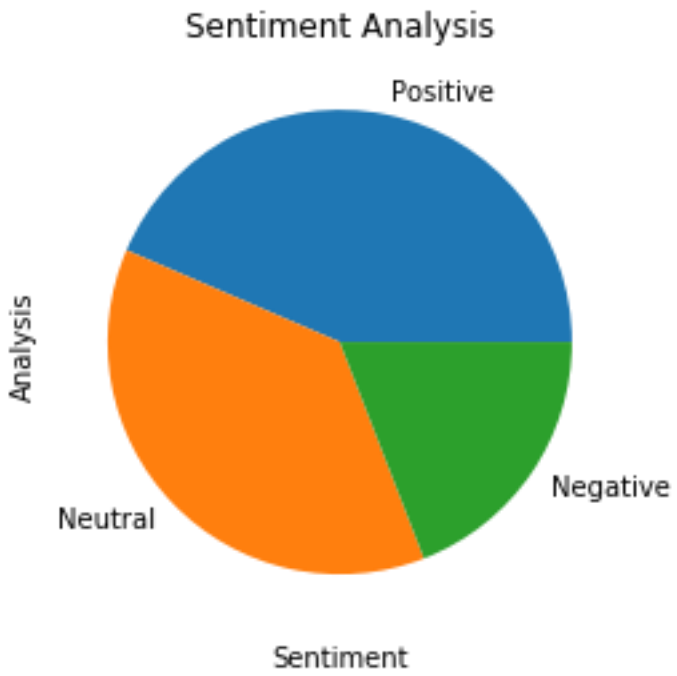

Fig 11: Analysis pie chart

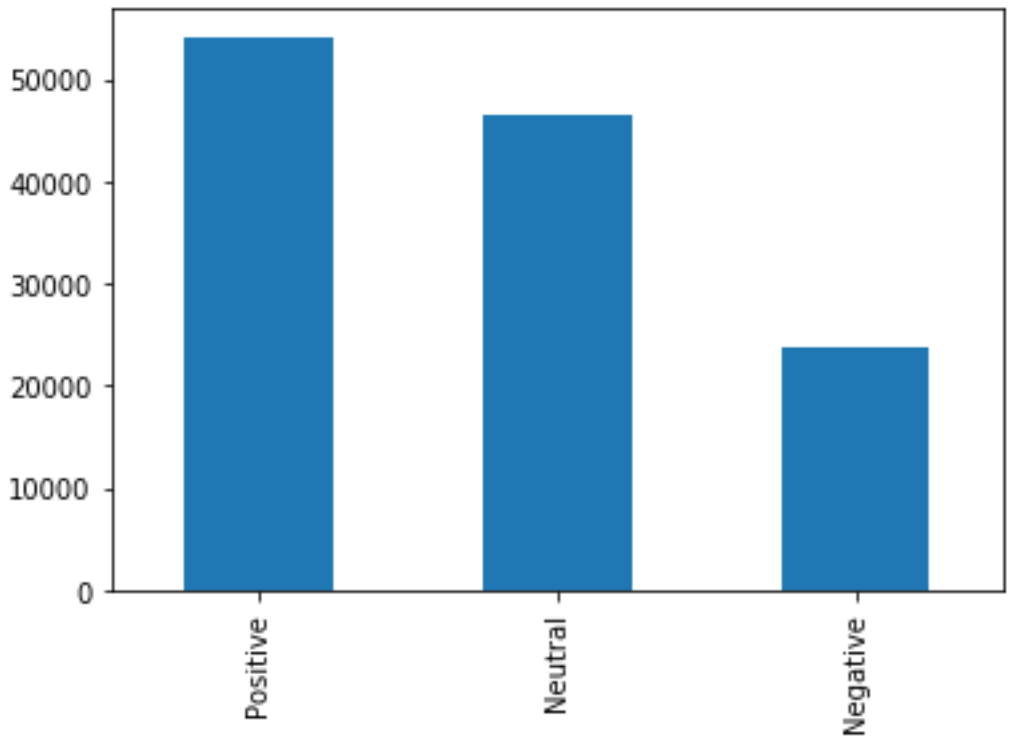

Fig 12: Analysis bar graph

3. Conclusion

Sentiment analysis is a field of study for analyzing opinions expressed in text in several social media sites. With this method, we reached to a conclusion that after three lockdowns in India, the **overall sentiment** of the people in India has been found to be **positive** as the number of positive tweets has been found to be the maximum, i.e., 54145.

```
Positive    54145
Neutral     46495
Negative    23744
```

References

[1] S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," *2019 International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, 2019, pp. 1-4, doi: 10.1109/ICCISci.2019.8716464.

[2] Priya Dwivedi [Blog], Available at: https://towardsdatascience.com/nlp-extracting-the-main-topics-from-your-dataset-using-lda-in-minutes-21486f5aa925

[3] https://spotle.ai/learn/NLP