IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006154|

Twitter Dataset Analysis for Covid-19 Pneumonia based on their Comments

Hitesh G[*1], Madhu HK [*2]

Dept. of MCA, Bangalore Institute of Technology, Bengaluru, India[*1]

Dept. of MCA, Assistant Professor, Bangalore Institute of Technology, Bengaluru, India[*2]

ABSTRACT: The news about Covid-19 spread like a wildfire on social media like twitter,facebook, Instagram etc. In a way awareness about the pandemic is good, But there is lot of misconception and unhealthy arguments going on in Social Media.

So to analyse about the opinion of people about Covid-19, it is ideal and necessary to conduct an Ex-post Facto evaluation of pandemic via virtual entertainment.

This includes perceptions of individuals who connect and share virtual entertainment on Twitter. So as a starting point of our survey, we present-now COVIDSENTI, a vast 90,000 tweets about opinion on Covid-19. From Feb to Mar 2020, It is further classified into Positive, Negative and Neutral comments. We used many combinations of highlights and classifiers to analyse the acquired tweets for opinion classifications.

For eg:- We noticed people who showed interest toward lockdown in Feb, then by mid March people were feeling moved.

I. INTRODUCTION

Covid is an original viral infection that has been recognized since it first appeared. The virus has spread to almost all countries and the effort to control the spread of it is being made in each one, The organization named world health organization, is working to prevent it, recognized it a pandemic on January 30, 2020. Antibody improvement is eagerly anticipated and demonstrated great commitment. There is a lack of scholarly focus on the subject to assist scientists. This obstructs study findings on the effects of COVID-19 on psychological well-being, as well as investigations into the global economic implications.

Because of the rise in atypical paranoid worries related to COVID-19, virtual entertainment stages like Twitter, Fb, Insta has been successfully dealing with fact-checking and reality-checking to conflict the spread of lie. Deception is stated as a calculated attempt to throw or deceive the general people with fake data. The necessitates the development of analytical methodologies that can be promptly communicated in order to interpret data flows and understand how mass opinion develops in pandemic scenarios.

Examining content posted on platforms like Twitter and Fb is a well-known method of detecting human profound articulation. Fright, numbers, realities, and the common concerns of people in general, of course, pervade the online entertainment arena, and this data, when analysed, can reveal a lot about the general state of mind and personality of the human population. The field of natural language processing (NLP) and its application to the study of web-based entertainment has grown at an exponential rate. Although, the oppositions of determining a text's characteristic relevance using NLP-strategies are still dangerous. Even the most recent advances in NLP have been found to be "powerless against adversarial texts" [42], [43]. As a result, it's critical to cultivate an understanding of the constraints of text categorization strategies, as well as known AI (ML) calculations.

Consider the following tweets1, in which clients express their thoughts about COVID-19. 1) Best wishes for the coming year. Hope this year bring you nice time, a huge amount of cheese, the resistance of Covid. 2) Now is the time to spread the word about the Corona Virus 200 affected people! I'm in a terrible mood. 3) What are the negative effects of Covid, and how widespread is it? 4) Experts warn that Covid is even more dangerous than Sars. 5) Mr Mohith was slain by the Covid. The models above depict the opinions and feelings of Twitter users who are interested in COVID-19.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006154|

II. RELATED WORKS

This piece is illuminated by numerous sources spanning various scholarly fields, and therefore, in this segment, the writing audit for feeling examination and literary investigation, also as ML techniques, Twitter and NLP, is presented. Similar information difficulties are arising and must be spoken, and crucial data qualities for information reconstruction, as well as machine learning methodologies, are critical tools [46]. Text-based inquiry handles the inspiration and examination of characters, text perceptions, semantics, and syntax, as well as linked exogenous and endogenous highlights of these devices.

In other work, an examination of over 70000 tweets sent over the course of a year was used to analyse client criticism by some company [38]. Because of the large amount of data gathered and examined by an idle designation calculation driven by recurrence based filtering procedures, fascinating bits of knowledge went unnoticed. Negative binomial and Poisson models have been used to investigate tweet notoriety [45]. In that review, the connection between points is also evaluated. There are seven disparity measurements used. Kullback-Leibler and Euclidean distances are shown to be the most effective in differentiating useful client-based intelligent technique associated locations.

Table-1
Keywords used to collect tweets

Keywords used to collect tweets
Coronoavirus
Corona
Covid
COVID-19
Pandemic
Corona Outbreak
Coronavirus Outbreak
Stay home
Coronavirus pandemic
Lockdown
Social distance
Quarantine

Philosophy:

COVID-Senti Data Collection and Labeling A. We included 90000 relevant tweets from 70000 clients that fulfilled the determination models out of roughly 21 lakh tweets slithered from Feb to Mar 2020. Our investigation uncovered 12 topics, including quarantine and staying at home. COVID SENTI was sub-divided into three sub-data sets: COVID-SENTI-A, COVID-SENTI-B, and COVID-SENTI-C, which represented positive, negative, and neutral opinions, respectively. Tweepy, an official Python Twtr API module, was used to compile the informative index (tweets). 1) Criteria for selection: COVIDSENTI has 2 months' worths of tweets. Our research was limit to English-language tweets. The watchwords used gave a printed corpus that was laser concentrated on subject.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006154|

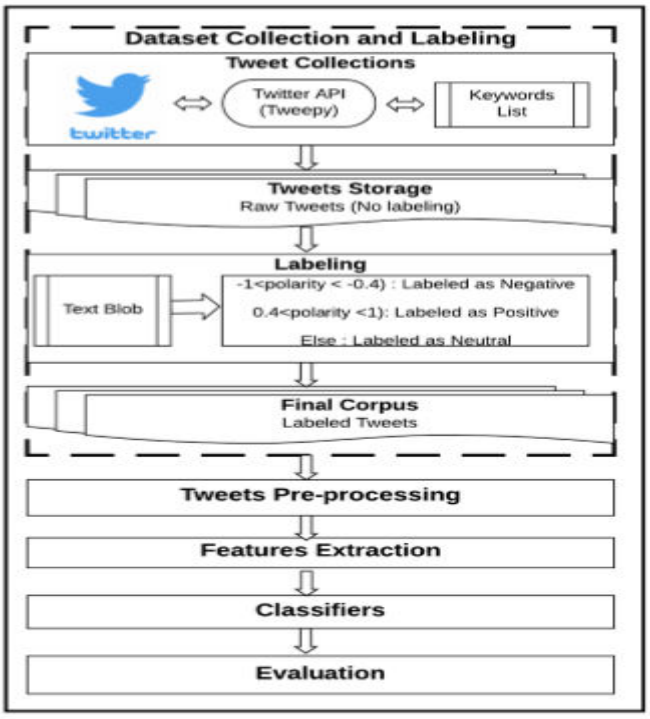

Fig. 2. Outline of the proposed system.

Coronavirus and its ramifications. Table I lists the watchwords that were used to gather tweets. 2) Identification: We used few guidelines to categorize each tweet as certain, negative, or nonpartisan in order to clarify the informational index. The TextBlob tool3 was used to categorize the close-to-home viewpoint into three categories: favorable, negative, and unbiased. TextBlob can demonstrate the demeanor of a sentence by analysing out the scores as an extremity minus one to one. If a comment's extremity is below 0.4, the sentiment is considered pessimistic. The tweet is considered favorable when it is more notable than 0.4. The polarity of a neutraltweet is between -0.4 and 0.4. Algorithm 1 provides the pseudocodes for tweet marking and emotion.

LTi=

Negative, P<-0.4 Positive, P>0.4 neutral, −0.4≥ P ≤0.4

(1)

where Pi is the extremity of Ti and −1 < P < 1.

Example of labeled tweets
• Happy New Year. May the Year of the Rat bring you good fortune, cheese in abundance and immunity to the coronavirus *(Labeled as "Positive")*
• Watching breaking news about the Corona Virus 200 infecteds now! Very sad *(Labeled as "Negative")*
• What are symptoms of coronavirus and where has it spread? *(Labeled as "Neutral")*

Fig. 3. Example of labelled tweet in Covid-senti

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006154|

Algorithm 1 Tweets Labeling Steps

Input: *Unlabeled Tweet*: T_u,
Output: *Labeled Tweet*: T_l,
Compute:
Positive: $T_{pos} = []$;
Negative: $T_{neg} = []$;
Neutral: $T_{neu} = []$;
Steps:
for t in T(t) do:
If (t is English):
Perform Labeling: (TextBlob)
If (-1 < polarity of t < -0.4):
Labeled as *Negative*
If (0.4 < polarity of t < 1):
Labeled as *Positive*
else:
Labeled as *Neutral*
Perform Pre-processing:
Remove punctuation, stop-words,
stemming, and lower-case all words
else:
Delete t
end for
Output:
Pre-processed labeled Tweets: $\mathbf{T}_l = [T_{pos}, T_{neg}, T_{neu}]$

Table 2
Data set distribution

Dataset\Label	Positive	Negative	Neutral	Total
COVIDSenti-A	1,968	5,083	22,949	30,000
COVIDSenti-B	2,033	5,471	22,496	30,000
COVIDSenti-C	2,279	5,781	21,940	30,000
COVIDSenti	6,280	16,335	67,835	90,000

The informational index, as said earlier, has 90000 tweets. For assessment and speculation reasons, it is also divided into 3 equivalent estimated subsets termed Covidsenti-a,Covidsenti-b, and Covidsenti-c in each opinion class. Table II describes the sharing of tweets. COVIDSENTI-A contains the great bulk of tweets in relation to government efforts to combat COVID-19. @Username, for example, says, "By all means, I have no belief in our administration." I get my information about covid obtained from off-the-shelf sources 4" Covisdsenti-b is a collection of tweets in-relation to Corona emergency, society exclusion, lockdown, and stay at home. As a result, it mostly covers the transient changes in people's actions as a result of the volume of cases, alert prompting data, and so on. "Covid goes illustrative," for example. Shut down everything and stay at home, China. Covidsenti-c is a group of comments related to Corona virus instances, episodes, and stays at home. As a result, it essentially displays patterns of human behavior in reaction to rise in number of incidents.

B. Pre-processing

Because tweets are frequently brief, unstructured, casual, and loud, the first step in analyzing an opinion is to preprocess the data [34]. To do this, the preceding set of methods is used in conjunction with the specified request to operate on the text. 1) Hashtags are used on almost every social media site to address issues. In general, hashtag's are meaningless to public opinion and can have an impact on the exhibition. As a result, in our first step, we cleaned up the test by deleting all unwanted hashtags. 2) The text is then case-overlayed as a next step. We crease all uppercase letters to bring them closer together in order to avoid perceiving a similar word as an alternate term due to upper casing. case in point 3) There are numerous terms that are all connected, especially hashtagged words, such as "stayhomestaysafe" and "coronavirus," which should be "remain at home stay safe" and "Covid," respectively. Following that, we do word division to achieve our goal. 4) Removing stop words is a well-known approach for reducing turbulence in literary information

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006154|

(a) (b) (c)

III. CONVERSATION AND CONCLUSION

Table-3
Top frequent words:

Keywords	Count
Corona	87,661
Coronavirus	78,459
Covid 19	26,239
Coronavirus cases	16,638
Coronavirus outbreak	7,419
Social distance	5,768
Positive coronavirus	3,413
Coronavirus pandemic	3,125
Coronavirus crises	1,981
Stay home	1,181

5) Lemmatization is the fifth phase, which involves morphologically examining words and returning them to their base form. We used nltk's given approach and assumption words (changing various types of words to its structure, example, "infections" to "infection" or "went" to "go"). 6) The first step in the inquiry was to remove hyperlinks, @mentions, and accentuation from the text. We exclude unusual characters, accents, and numerals from the informational collection because they don't help us recognize feelings.

C. Analyze exploratory We use exploratory investigation in this phase to gain a more full picture of our informational index. 1) Keyword Trend Analysis: We began by looking for catchy patterns in our preprocessed corpus to determine the most often referred terms. We noticed the beings are discussing about Covid, the Covid event, social separation, the Covid Panddemic, Covid emergencies, and staying in home. Measurements on the TOP10 with frequencies used watchwords are obtained, and results are seen in Table III

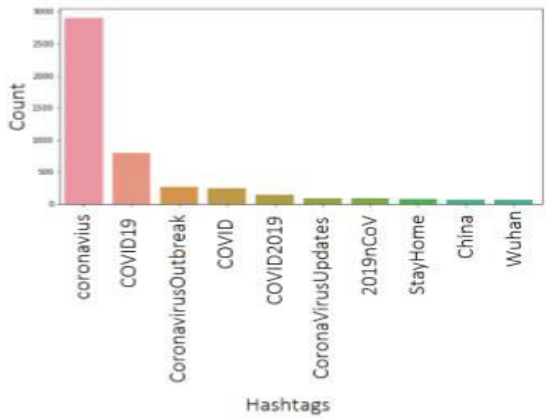

Fig 5. Top hashtags in the COVID-Senti informational index.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006154|

LDA is a subject displaying calculation, which implies that a text is composed of a series of points. After learning the LDA, subjects depicted by word dispersion and document point distribution are learned. The number of subjects in LDA is set at six. After LDA preparation, the subjects addressed by a dispersion of words and the point disseminations of the records are learned. Figures 6 and 7 show, respectively, the word haze of terms within the top six subjects and the dispersion of the top-six predominant points in the corpora using LDA

The above img is about 6 topics

Table-4

Topic 1:	coronavirus	cases	new	says	confirmed	china	cruise	ship	coronavirus cases	covid
Topic 2:	coronavirus	china	death	novel	toll	due	novel coronavirus	death toll	wuhan	see
Topic 3:	coronavirus	world	like	wuhan	case	first	spread	wuhan coronavirus	south	emergency
Topic 4:	coronavirus	get	fears	know	hospital	coronavirus fears	impact	need	test	patients
Topic 5:	coronavirus	outbreak	china	coronavirus outbreak	trump	global	news	chinese	amid	live
Topic 6:	virus	corona	corona virus	coronavirus	people	flu	mask	got	think	one

before the middle of March 2020. After mid-March 2020, the number and severity of negative feelings decreased and moved as people became more receptive to specialist-implemented lockup and social separation settings.

D. Highlight Extraction

Include extraction is accomplished using vectorization techniques and word bedding in this study. For vectorization, term recurrence reverse record recurrence (TF-IDF) was used. Pretrained Word2Vec, GloVe, and fastText embeddings prepared Wikipedia with 300-D vectors are also used for word embeddings. In addition, for Twitter opinion research, we used crossover models like as half breed positioning, which absorb the feeling and factors of ttweets.

This explains the trial stage that was used to evaluate the presentation using benchmark informative indexes and provide benchmarked outcomes to the correlation motivation. We used precision and a ten-fold cross-approval process. Different ML, DL, and half-and-half techniques are used to account for the usual outcomes. Tables V-IX summarize the findings, which are displayed in Fig. 9. We used traditional techniques like TF-IDF,word insertion based models like, half breed models like IWV and HyRank, like bert, DistilBERT, xlnet, and ALBERT to build down the baselines for ML classifiers. The consequences of tf-idf based classification are as follows:

TABLE-6

COMPARISONOF MLWITHWORD EMBEDDINGS

Machine Learning Classifiers with Word Embeddings.					
ModelsDatasets		COVIDSenti-A	COVIDSenti-B	COVIDSenti-C	COVIDSenti
Word2Vec	RF	0.764	0.732	0.753	0.769
	DT	0.762	0.745	0.741	0.766
GloVe	RF	0.716	0.702	0.722	0.726
	DT	0.693	0.689	0.694	0.701
FastText	SVM	0.801	0.792	0.783	0.815
	NB	0.732	0.745	0.721	0.735
	RF	0.823	0.841	0.802	0.845

TABLE-7

COMPARISONOF DL CLASSIFIERSWITHWORD Embeddings

Deep Learning Classifiers with Word Embeddings.					
Models\ Datasets		COVIDSenti-A	COVIDSenti-B	COVIDSenti-C	COVIDSenti
Word2Vec	BiLSTM	0.765	0.745	0.749	0.769
GloVe	BiLSTM	0.768	0.729	0.743	0.771
DCNN- (Glove+ CNN)		0.834	0.832	0.864	0.869

Table-8

Comparision Of Hybrid model

Hybrid Model				
Models\Datasets	COVIDSenti-A	COVIDSenti-B	COVIDSenti-C	COVIDSenti
IWV	0.763	0.749	0.735	0.771
HyRank	0.854	0.865	0.877	0.881

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006154|

Table-9
COMPARISON OF Transformer-BASEDLMS

Fine-tuning of Transformer based language models				
Models\Dataset	COVIDSenti-A	COVIDSenti-B	COVIDSenti-C	COVIDSenti
distilBERT	0.937	0.929	0.926	0.939
BERT	0.941	0.937	0.932	0.948
XLNET	0.924	0.914	0.920	0.933
ALBERT	0.914	0.920	0.910	0.929

IV. CONCLUSION

Web-based entertainment has been widely used both for and against deception and confusion since the explosion of COVID-19 crazy concepts. The topic of twtr feeling on Covid-19 related posts is addressed in this article. In the investigation of COVID-19-related opinion, we compare feeling examination methodologies. Our findings show that in February, the people favoured the lockdown and stay-at-home request; but, by first half of march , their views had shifted. The cause for change in perspective is unclear, but it could be because of incorrect information being shared via internet entertainment; as a result, there is a need to nurture a active and lean general health presence to combine the propagation of false news. To aid researching in the society, we made a massive COVID-19 benchmark opinion examination informational collection openly available.

REFERENCES

[1] C. C. Aggarwal and C. K. Reddy,Data Clustering: Algorithms and Applications. Boca Raton, FL, USA: CRC Press, 2013.
[2] N. Ahmad and J. Siddique, “Personality assessment using Twitter tweets,” Procedia Comput. Sci., vol. 112, pp. 1964–1973, Sep. 2017.
[3] T. Ahmad, A. Ramsay, and H. Ahmed, “Detecting emotions in English and Arabic tweets,” Information, vol. 10, no. 3, p. 98, Mar. 2019.
[4] A. Bandi and A. Fellah, “Socio-analyzer: A sentiment analysis using social media data,” in Proc. 28th Int. Conf. Softw. Eng. Data Eng., in EPiC Series in Computing, vol. 64, F. Harris, S. Dascalu, S. Sharma, and R. Wu, Eds. Amsterdam, The Netherlands: EasyChair, 2019, pp. 61–67.
[5] F. Barbieri and H. Saggion, “Automatic detection of irony and humour in Twitter,” in Proc. ICCC, 2014, pp. 155–162.
[6] R. Bhat, V. K. Singh, N. Naik, C. R. Kamath, P. Mulimani, and N. Kulkarni, “COVID 2019 outbreak: The disappointment in Indian teachers,” Asian J. Psychiatry, vol. 50, Apr. 2020, Art. no. 102047.
[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Jan. 2003.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details