International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 1, ISSN No. 2455-2143, Pages 256-262
Published Online May 2020 in IJEAST (http://www.ijeast.com)

STUDY AND ANALYSIS OF MISINFORMATION SPREAD DURING THE COVID-19 IN INDIA

Mohd Saquib
Dept. of Electronics & Instrumentation
Galgotias College of Engineering & Technology
Greater Noida, Uttar Pradesh, India

Mohd Kashif
Independent Researcher
Noida, Uttar Pradesh, India

Abstract—**The aim of this study was to analyze the trend and evaluate the impact on misinformation spread with an increase in the number of positive cases for novel coronavirus diseases (COVID-19) in India using Text Analytics and Natural Language Processing(NLP). About 726 unique misinformation were scraped from various IFCN certified fact-checkers platforms during the period of Jan 15, 2020, to April 20, 2020. Each story has systematically annotated Headline, Date, and the Domain. The category for each story was also defined using the Latent Dirichlet Algorithm(LDA). We use Bar-Graphs to plot trends of misinformation, Word Cloud to visualize the important tags in the story. After critical analysis we obtain the results that there was a significant conspiracy theory related component in the earlier stories, as the coronavirus disease (COVID-19) spread increased in India in the early stage, the story trend changes to cure and prevention from coronavirus disease(COVID-19), then after the declaration of lockdown in India the story trend changes to business and economy, Later in last week of March, religious and cultural references appeared in huge number.**

Keyword— **Misinformation, Latent Dirichlet Allocation, COVID-19, Natural Language Processing.**

I. INTRODUCTION

The internet has become an important source of health information for users worldwide [10]. Along with correct information, internet allows for instant access to, and dissemination of, misinformation around the globe [2]. With the emergence of novel coronavirus disease (COVID-19) from its epicenter in Hubei, China. The misinformation on various social platforms and on internet erupted like a wildfire, the misinformation and the positive cases of COVID-19 has shown the linear relationship. With the increase in a number of positive cases in India, there is a rise in the number of debunked misinformation, especially following the third week of March 2020. The momentum of misinformation had already stated a huge rise before Prime Minister Narendra Modi announced the Janta Curfew on 22nd March 2020, from then the increase in misinformation has been consistently increasing. At the Munich Security Conference on Feb 15, the WHO Director-General said: "We're not just fighting an epidemic; we're fighting an infodemic" [3]. We also found out that in the historical cases of an epidemic like Ebola 2014-15, World Health Organization (WHO) noted rumors circulating on the internet claiming that certain products or practices could prevent or cure Ebola virus disease [4]. After COVID-19 was declared a Public Health Emergency of International Concern, WHO's launched a new information platform called WHO Information Network for Epidemics(EPI-WIN), to share information with groups [5]. It is important to verify the information on the internet to prevent the panic and misinformation associated with COVID-19. The fake news spread faster than the virus, the internet is the main source of information in India; currently over 560 million internet users, India is the second-largest online market in the world [6]. Online health information has become popular among non-health personnel users, for users with non-medical background, it is difficult to judge the credibility of health information on the internet. The Internet can serve as a valuable source of information as well as misinformation about the virus globally which can lead to panic situations and creating a so-called infodemic [3]. Propaganda and persuasion methods are widely used on the Internet easily reaching target groups, promoting conspiracy theories, and polarizing societies based on religion, politics, and community. The panic-related behavior due to a virus outbreak is influencing the infodemic spread [7]. This could lead to destructive behavior, such as xenophobia against people from affected countries, community, or religion [8]. The internet activities are being analyzed worldwide to better understand the perception and types of misinformation spread at a specific time. This helps in a way to be prepared with better guidelines and precautions to combat misinformation spread and destructive behavior of people [9]. The internet is good for the propagation of views often contradicting medical science. The Internet can serve as a valuable source of information as well as misinformation about the virus globally. Therefore, the need for critical analysis of misinformation i.e. the trend and categorization of misinformation with respect to increasing in the number of cases of COVID-19 and particular events like lockdown in India have been developed.

II. METHODOLOGY

We scraped 726 unique misinformation instances were scraped from various IFCN certified fact-checkers: BOOMlive, Factly, Indiatoday Fack Check, Quint Webqoof, and NewsMobile Fact Checker during the period of Jan 20, 2020, to April 20, 2020. Each misinformation is considered as a data-point with the following attributes - Date, Domain, and Headline of the story. Here the Date is defined as the date at which the story was updated on the fact-checker platform, Domain is the platform from where the story is taken, and the Headline is the Title of the Misinformation. This information of attributes we obtained systematically while scraping the data from these platforms. We also categorized the misinformation into the following 7 categories Table 1, To define these categories we have used Latent Dirichlet Allocation (LDA) [10], which is a probabilistic generative model for collection of discrete data such as text corpus. LDA is widely used to classify text in a document given. We used LDA on our stories and classified the misinformation into 7 categories as shown in Table 1, but before using LDA we have to first preprocess our data.

Table 1 : Categories of Misinformation

S.No	**CATEGORIES OF MISINFORMATION**	
	Category	***Description***
1	Conspiracy	Message that is related to any conspiracy theroy e.g. theories suggesting COVID-19 a bio-weapon or a 5G effect.
2	Casualty	Message that is related to illness, no of cases, sufferings, panic and doctored statistics.
3	Business and Economy	Message that is related to scams, panic-buying, offers, hoax of shortage of essential commodoties and target business.
4	Cure and Remedy	Messages that are related to cure, remedy for the COVID-19, or remedies of prevention from virus
5	Environment	Messages that are related to plants, animal etc.
6	Political	Messgae related to government notifications, orders, politician and politics.
7	Culture	Message relaed to culture, religion or community

A. Pre-processing:

To use LDA for categorization we first Pre-processed our misinformation documents which consist of following steps-

1. Tokenization: Splitting the text into sentences and further sentences into words. Lowercase the words and remove punctuation. Words that have less than 3 characters were removed.
2. Stopwords were removed
3. Lemmatization: Words in the third person are changed to first person and past and future tense is changed into the present.
4. Stemming: Words are reduced to their root form. After Pre-processing, we created a dictionary for the processed document containing the number of times a word appears in the training set using the Bag of Words technique [11]. We run our LDA model using the genism. Model. LdaMulticore [12] after using Bag of Words. As a result, for each story, we get the words occurring in that topic along with its relative weight. Using weight we classified the story into the above mention 7 categories as shown in Table 1.

After defining the categories, we analyzed the trend for these categories with respect to time the misinformation spread. To analyze the trend of misinformation categories in India, we selected different time duration in which important events, orders from the government, for e.g. lockdown notification. as shown in Table 2.

Table 2 :Time Period Selected

S.No	**TIME PERIOD SELECTED FOR ANALYSIS**	
	Time Window	***Description***
1	1 Jan - 1 Feb	In this period the cases for COVID-19 were quite low, first positive case was reported on 30 Jan [14].
2	11 March - 21 March	On 11 Mar, WHO Declares COVID-19 a pandemic.
3	17 March - 22 March	On 19 March, PM Modi addresses to nation and announced a 14-hours voluntary curfew on 22 March 2020 [14].
4	22 March - 26 March	On 24 Mar, the prime minister ordered a nation wide lockdown for 21 days [14].

S.No	TIME PERIOD SELECTED FOR ANALYSIS	
	Time Window	*Description*
5	29 March - 10 April	On 30 Mar, first hotspot was found in India.

C. Plotting:

The analysis of the trend of misinformation for these time window was done as in Table 1, using plotting tools available in Matplotlib, which is a comprehensive library for creating interactive visualization in Python [13]. The normalization of the plot was also done to adjust values and irregularity in a number of samples of information for a particular time period.

D. Word Cloud:

We also generated the word clouds for the stories present in a particular time window and analyze the type of word and phrases occurred in that period. Word cloud. A word cloud is an image made of words that resembles a shape like a cloud. The size of the word shows how important the particular word in that story.

III. Results

A. 10 Jan – 10 Feb :

In the period of January to February, the positive cases of COVID-19 were very low in India, the first positive case in India was observed on 30 January [14]. The trend in categories of misinformation is shown in from Fig 1, we can visualize that during the early spread of COVID-19, most of the misinformation spread were supporting conspiracy followed by casualty and cure. At this period of time, a lot of misinformation spread for supporting some conspiracy theory against any country or community. The word cloud for the period of Jan-Feb is shown in Fig 2, from where we can observe most of the important terms used in stories were supporting some conspiracy theory.

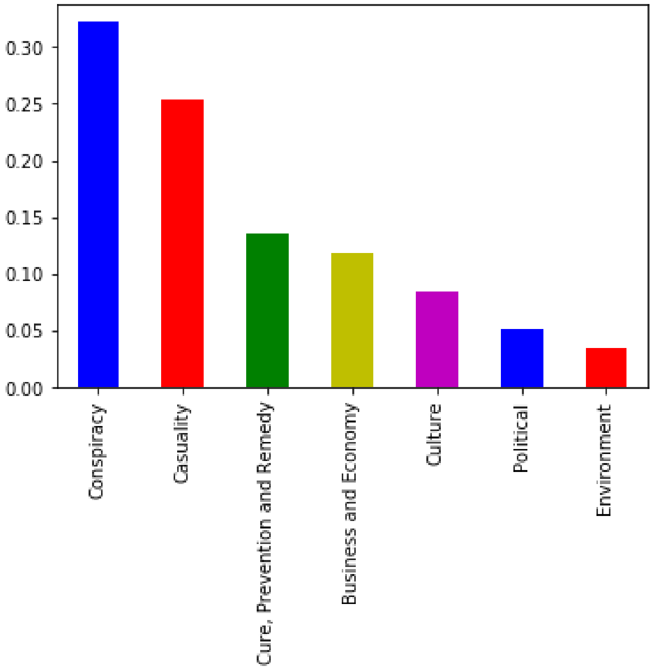

Fig. 1.-Category trend (Jan-Feb)

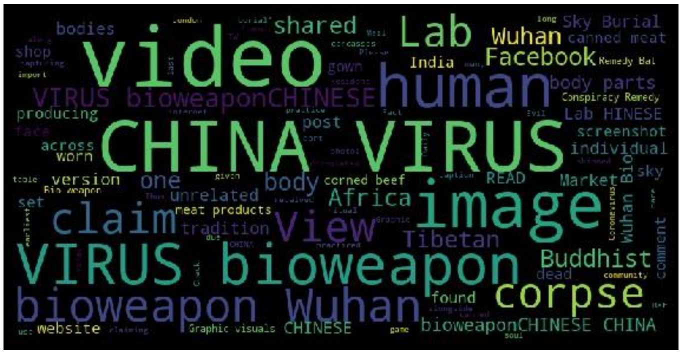

Fig. 2- Word Cloud (Jan-Feb)

B. 10 Jan – 10 Feb:

During the period of 11 March to 21 March. The positive cases in India started increasing slightly, On 11 March World Health Organization (WHO) Declares COVID-19 a pandemic. The confirmed positive cases of COVID-19 was 194. The trend for this period is shown in Fig. 3, from the figure we analyzed that after the slight increase in positive cases of COVID-19 in India, and after World Health Organization (WHO) declaration of COVID-19 being a pandemic, the misinformation trends changes to Cure, Prevention and Remedy- the messages suggesting home remedies, preventive measures from the COVID-19, and vaccine-related disinformation started to spread very rapidly. Also, from the word cloud as shown in Fig. 4, we see distinct trends of suggesting the treatment of COVID-19 with home-remedies, terms of some herbs commonly used in India can also be seen in our word cloud. The category of misinformation for environment, conspiracy and culture were not so significant during the time duration of 11 March to 21 March, however the spread of political and casuality stories seems to be equally spread during this time period.

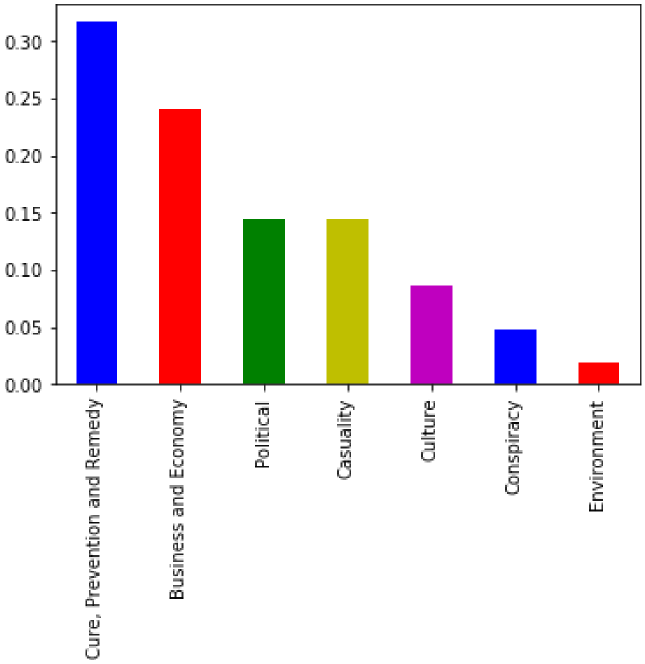

Fig. 3. Category trend (11 March – 21 March)

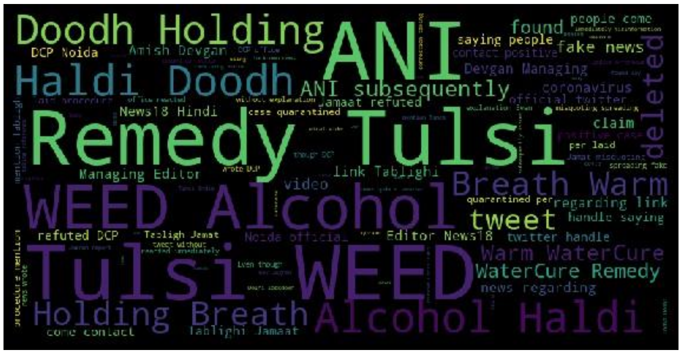

Fig. 4. Word Cloud (11 March - 21 March)

C. 17 March – 22 March :

During the period of 17 March to 22 March, the positive cases count increases to 300. On 19 March, PM addressed to the nation and announced a 14-hour voluntary public curfew on 22 March 2020, from that we can visualize as shown in Fig. 5, that the misinformation from both category Cure and Business & Economy was at peak. Also, from the word cloud as shown in Fig. 6, we can see words like 'Essential Services', 'Lockdown', 'Services', 'Curfew' , 'Remedy' were the most important tags of stories from this period.

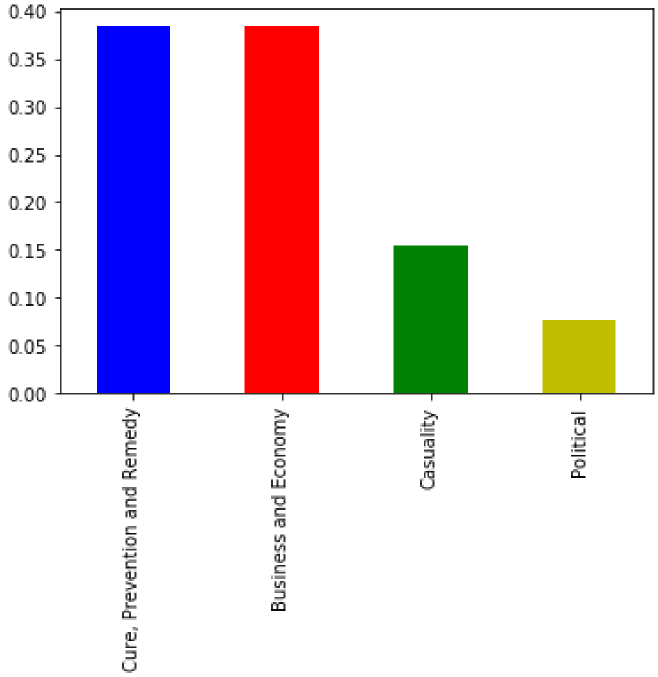

Fig. 5. Category trend (17 March - 22 March)

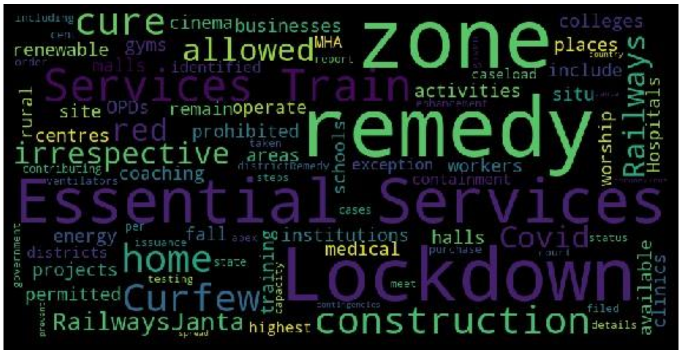

Fig. 6. Word Cloud (17 March–22 March)

D. 22 March – 26 March :

During the period of 22 March to 26 March, the number of positive cases in India was above 500, On 24 March prime minister ordered a for the period of 14 days, from the Fig.7, we can analyze that in this period misinformation category- Cure, Business and Casualty have shown an equal peak. We observed in this period a large share of the casualty-related messages was driven by the increase in casualty in Italy. Also, misinformation underlying in Business and Economy category can be seen from the word cloud, where words like 'Shortage', 'Lockdown', Panic-Buying, etc were the most important terms in the stories from this period. Also, it can be observed from the bar-plot misinformation categories: Political, Culture, Conspiracy, and Environment have shown a significant drop in percentage during this period of time.

In the word cloud, terms related to these terms were found very less, they were more terms related to Casualty- terms like ‘Died’, ‘Attack’ can also be seen as the important term used in the misinformation spread during this period of time as shown in the word cloud.

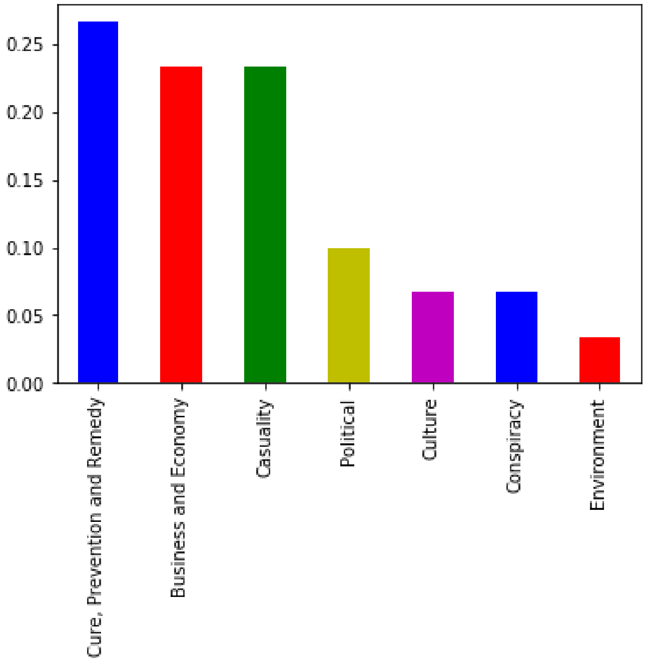

Fig. 7. Category trend (22 March - 26 March)

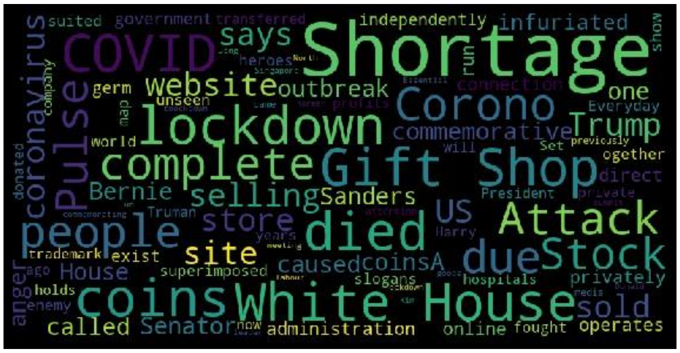

Fig. 8. Word Cloud(22 March - 26 March)

E. 29 March – 10 April:

During the period of 29 March to 10 April, as the visualization shown in Fig. 9, shows the culture-related misinformation was at the peak during this period followed by the category of Business and Economy, Political and Casualty, from the word cloud as shown in Fig. 10, it can be seen, culture-related misinformation has a very significant culture-related growth during this time period in India. It can be observed that there were a lot of conspiracy stories in the earlier stories, as the COVID-19 started to spread in India, but later the religious and cultural misinformation appeared an increasing number from the last 10 days of March. It can also be seen at this time the category of Business and Economics was also at peak followed by the Political category, the Environment, Conspiracy, and The Cure, Prevention and Remedy category has shown a significant drop during this time period.

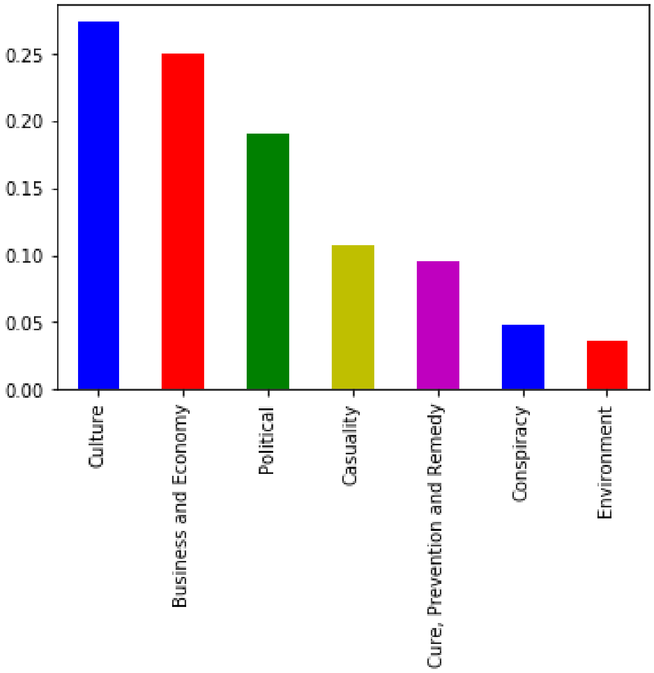

Fig. 9. Category Trend (29 March - 10 April)

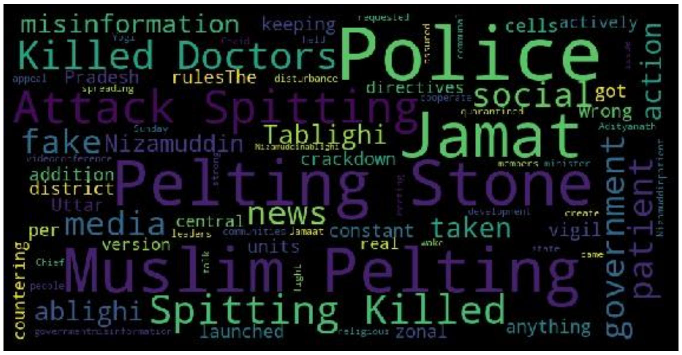

Fig. 10. Word Cloud(29 March - 10 April)

F. Overall Trend from (January - April):

In the overall trend of misinformation, it can be seen from the bar plot in Fig. 11, the misinformation related to Culture, Business and Economy, Casualty, Political, and Cure, Prevention and Remedy have an equal share of misinformation with a very little difference, Cultural misinformation being at the top, which means that these type of misinformation were spread equally but at the different time period. Also, from the word cloud, we can see the mix of the terminology of words been used in stories during this period of time, words related to Cure, Casualty, and Cultural were mostly found during this period of time as shown in the word cloud Fig. 12. The plot used is normalized, and the y-axis of the plot represents the percentage. We also observed that misinformation related to the environment was low in this period, also conspiracy misinformation in beginning was at a peak, but it decreases up to a great extent after the authorities started a campaign to spread accurate information in societies. The overall trend gives a clear idea about the category of misinformation, which category to be considered as important and which are not so important. The environment-related misinformation was low in the overall period of January to April.

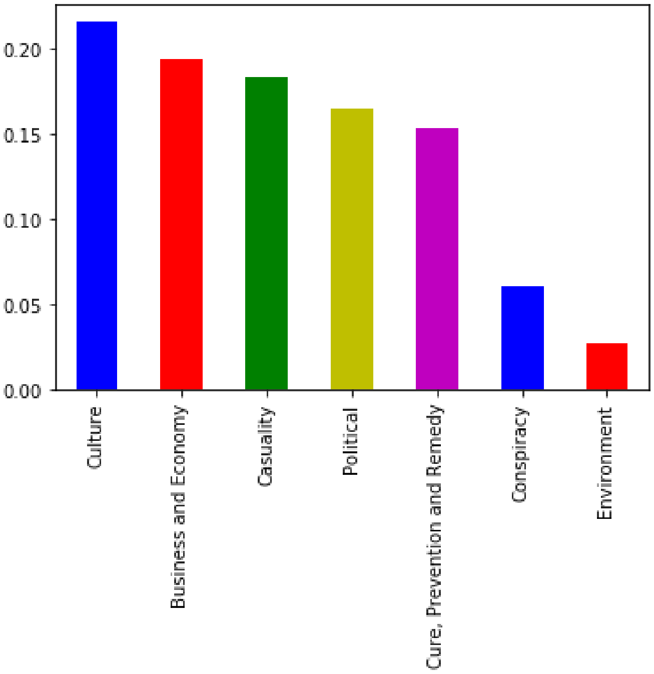

Fig. 11. Category Trend (Overall)

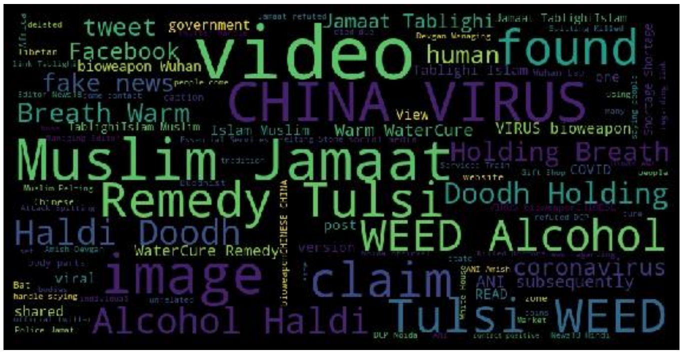

Fig. 12. Word Cloud(Overall)

IV. CONCLUSION

From the results, it can be concluded that in India, the misinformations has shown varied trend and changes with respect to different time duration and an increase in the number of cases of COVID-19 also have the adverse effect on the spread of misinformation. In the starting when the COVID-19 cases in India were low, it is observed that most of the misinformation was related to a conspiracy theory, but later as cases grew and certain orders came from the government the trend of misinformation changes to cure and remedy for COVID-19, On notification of lockdown in India, the misinformation category have shown a change to business and economy category. Later from March 20, there was huge exponential growth in culture types of misinformation. It was also observed that preventive measures from authorities to stop the misinformation have a very successful impact on the perception of the public regarding the pandemic so it can be very helpful for the authorities to fight the infodemic if they understand the trend and categories of misinformation being spread.

V. REFERENCES

[1] Hagg E., Dahinten VS., and Currie LM.(2018). The emerging use of social media for health-related purposes in low and middle-income countries: A scoping review, in International Journal of Medical Informatics,(pp.92 - 105).

[2] Granik V., and Mesyura.(2017) Fake news detection using naive Bayes classifier in IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017,(pp. 900–903).

[3] Zarocastas J.,(2020). How to fight an infodemic, in The Lancet,(p676)

[4] Chakraborty, S., Pal, J., Chandra, P., Romero, D. (2018) Political Tweets and Mainstream News Impact in India: A Mixed Methods Investigation into Political Outreach. Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies (10)

[5] Isaac Chun-Hai Fung, King-Wa Fu, Chung-Hong Chan,, Benedict Shing Bun Chan, Chi-Ngai Cheung, Thomas Abraham, and Zion Tsz Ho Tse.(2016) Social Media's Initial Reaction to Information and Misinformation on Ebola - Facts and Rumors 2014 in Public Health Reports,(pp.461-473)

[6] Bhattacharjee B., Biswaas A., Agarwal R., and Mishra R.(2019) Internet in India in IMRB Intrenational and IAMAI.

[7] Matthew J., Isabelle B., Camilla H., and Altman D.,(2018).Assessing risk of bias in studies that evaluate health

care interventions: recommendations in the misinformation age in Journal of Clinical Epidemiology(pp.133-136)

[8] Rzymski P., and Nowicki M.,(2020) Preventing COVID-19 prejudice in academia-Science, 20 Mar 2020,(pp. 1313)

[9] Funke D., and Falmini D., Poynter Resources : A guide to anti-misinformation actions around the world in Poynter [ONLINE].

[10] Blei, David & Ng, Andrew & Jordan, Michael & Lafferty, John.(2003). Latent Drichlet Allocation in Journal of Machine Learning Research (2003),(pp.993-1022)

[11] Zhang, Yin & Jin, Rong & Zhou, Zhi-Hua. (2010). Understanding bag-of-words model: A statistical framework in International Journal of Machine Learning and Cybernetics,(pp.43-52)

[12] David M.,(2012) Probabilistic Topic Models in Communication of the ACM(Vol.55| No.4)

[13] Lowe D.,(2003) Distinctive image features from scale-invariant key points in International Journal of Computer Vision ,(pp.91-110)

[14] Minisry of Health and Family Welfare [ONLINE] .

[15] Erin L.,(2003) Informed Consent in the (Mis) Information Age in Journal of Obstetrics and Gynaecology Canada,(pp43-48)