An effective architecture for prediction of severity levels of anxiety,depression,stress during the dissemenation of Social media Infodemic at the times of COVID-19 using multiclass classifier

Mrs.Garima Joshi[1], Dr.Prashant Panse[2]
[1,2]*Computer Science and Engineering, Medicaps-University (AICTE Approved) Indore, India*

***Abstract*—During the COVID-19, an infectious disease caused by the SARS-CoV-2 virus that spread across the world ,a huge amount of population got affected. It is directly affected to the mental and physical health of every stage of age and hence thereby immensely affects the work –life balance either directly or indirectly. Almost one year was the life changer phase of everyone's life. The mental state was affected due to financial issues, relational issues, environmental issues and some constrainable issues related to prevention from COVID-19.Some sort of pressure of survival in the profession and personal life is the big challenge for everyone ,nowadays. During the lockdown period, frequent access of news channels, online and social media give birth to anxiety and stress .Increase in the media sources and variety of information presented on the single topic leads to develop multiple questions in our minds .Infodemic plays the vital role as the various information sources like news channels, social media, E- newspapers provides excessive information out of which some are essential for us but some information are un-essential and such un-important information has become the high risk for the mental health of individuals. In this research, we will attempt to analyze the various types of media content impacting mental health of the individual during COVID-19 therefore affecting the work life balance. As a part of this, we analyze different social media contents which are the part of infodemic and will apply data science techniques to analyze that the which part of information is misinformation in the terms of affecting mental state of anyone .**

***Keywords*—Epidemic, infodemic, misinformation, disinformation**

I. INTRODUCTION

Today is an era of competitive environment in personal and professional life, everyone is facing challenges for the individual upbringing. Despite, each of us facing it with great efforts .It does affect the mental and physical health. Infodemic is an integral part of day today life, thereby affecting the mental health of almost all the age group. There are various sources of infodemics which are accessible free of cost and available on the single click. The major challenges are to determine which of this information are essential and non-essential. Some of them, are misguiding in many aspects and hence affecting the growth of individual as it is directly proportional to the state of mind. Previous year is almost the life upheaving year for everyone. Upheaving in financial, professional, personal, political filed and related misinformation is responsible for disturbing mental health. Spreading of misinformation is a global issue , whether it can be through social media platform on any other source of media. During the epidemic, controlling the explosion of misinformation is one of the biggest challenges. Health mis-information plays vital parts which can be defined as the false information related to medical science and available through medical experts that has been accessible on almost all the popular social media platforms including YouTube, Facebook, Instagram, as epidemic is concerned , mis-information related to safety precautions against COVID-19,remedial medicine, vaccination, social distancing and many other myths and subjecting to medical science are considered to be high risk for the stability of mind. As multiple misinformation of a particular themes and its responses impacting to majority of people , globally. The term infodemic has been coined to outline the perils of misinformation phenomena during the management of virus outbreaks2 , since it could even speed up the epidemic process by influencing and fragmenting social response [1]. . The avalanche of human response is being facilitated by the flow of information from the broadcast world of traditional media but, in particular, by the networked world of social media. Social media is a significant conduit for news and information in the modern media environment, with one in three people in the world engaging in social media, and two thirds of those on the Internet using it [2]. The spread of misinformation can undermine public health strategies [96] and has potentially dangerous consequences [97], [98]. For example, online rumors accusing 5G deployments of causing COVID-19 led to mobile phone masts being attacked in the UK [99]. Wikipedia maintains an up-to-date list of misinformation surrounding COVID-19 [100]. This confirms the spread of a number of dangerous forms of misinformation, e.g., that vinegar is more effective than hand sanitizer against COVID-19. Naturally, users who believe such misinformation could proceed to undermine public health. One important use case would therefore be to develop classifiers and techniques to stem this flow.

II. LITERATURE SURVEY

In this paper, Malteo Cinelli et al performed a comparative analysis on five different social media platforms during the COVID-19 health emergency. They have worked on content consumption dynamics and assess users engagement and interested about the COVID-19 topic characteristics the evolution of the discourse over time. They analyzed the spreading of questionable information for all channels , finding that Gab, one of the those five different social media is the environment more susceptible to misinformation diffusion. They have provided COVID-19 rumors amplification parameter for social media platform. As the material and methods , the first part of their research is Data collected during specific time period and the sources of data collected is defined. After that , in the text analysis, extraction and analysis of all topics related to COVID-19 is carried out by applying Natural Language processing techniques to written content of each platform then to assess the topics around which the perception of COVID-19 debate is correlated. This article looks at the amount of conversation taking place on social media, specifically Twitter with respect to COVID-19.Based on the themes of discussions is emerging from and how much of it is connected to other high or low quality information on the internet through shared URL links .They analyzed a variety of sources including blogs, newspaper /media and medical organization to get a variety of examples. They identified common themes among the different myths. They searched for each of these myths in their tweet collection. They preprocessed the tweets to normalize capitalization and remove punctuation and URLs.[4]They have presented work on COVID-19 misinformation on Twitter .they have analyzed tweets that have been fact checked by using techniques common to social media analytics[5]Stinne giaadam et al aimed to explore different airculations of people’s understanding ,handling and evolution of (true) information ,misinformation and disinformation in general and specifically linked to social media ,related to the COVID-19 pandemics to illuminate the complexity of the construction of true information.[6]Participants in this study made decisions that resulted in coping behaviors and strategies to manage news overload. Thus, whilst all except one said they believed this was a time of information overload, and, indeed, all at some point avoided or ignored the news, they still continued to access and consume the news. Each individual participant’s management solutions to news overload will be determined by their natural disposition to news consumption; pre-COVID-19 media habits; motivations for consuming news; levels of media and digital literacy; self-efficacy; demographics; and personal factors. The different methods of managing news overload (MNO) have, in practice, interconnected and overlapped and helped participants to control their news diet and deal with potential, perceived and real occurrences of overload. Both the quantity and variety of news entries in the media diaries point to respondents satisficing.[7]They have categorized ideas into a set of five COVID-19 infodemic management areas ,using narrative analysis adapted to national contexts and practices .Responses to the COVID-19 pandemic and the related infodemic require swift, regular, systematic ,coordinated action from multiple sectors of society and government .They provided the framework for managing the COVID-19 infodemic-methods and results of an online by emphasis on the phrase “an infodemic cannot be eliminated but it can be managed”.[8]In this paper authors provided gives light on dissemination of misinformation on various social media platform during the spread of COVID-19. Social networks have become central to the rapid dissemination of scientific information and for administrative pandemic monitoring and control in the field of distance learning, remote monitoring and health care.[9]They have introduced the data driven methods requires that data is trustworthy, discoverable, accessible, reusable, and frequently updated. One of them is Data curation. Data curation is an active management of data over its life cycle to ensure it meets the necessary data quality requirements for its effective usage. Data curation process can be categorized into different activities such as content creation, selection, classification, transformation, validation, and preservation. A key trend for the curation of COVID-19. Data sets are different open-source communities like kaggle and github.[10]This article basically focused on information disorder and introduce a new conceptual framework for examining information disorder, identifying three different types of information i.e. misinformation, disinformation and mal information. Misinformation is when false information is shared , but no harm is meant. Dis-information is when false information is knowingly shared to cause harm. Mal information is when genuine information is shared to cause harm often by moving information designed to stay private into the public sphere.[11]In this paper, Soroush Visoughi et al performed the investigation on the differential diffusion of true and false new stories over twitter from 2006 to 2017 and classified them. After categorization they have found that diffusion of false information is more the true information. False information based on terrorism, natural calamity, finance and science were more diffused. In addition to that they said that the robots spread both true and false information at the same rate but human are more likely to spread the false information.[12]They have introduced an infodemic response checklist for promoting more efficient health communication strategies to alleviate the efforts of the current outbreak of mis-information and any other outbreaks of mis-information and any other outbreaks that may arise in the future. In this paper, they have emphasis on overcoming the COVID-19 infodemic by mentioning that surfeit of information about the COVID-19 pandemic spread widely. While some information was true much was false. This resulted in an “Infodemic” whereby waves of misinformation and rumours on the pandemic interfered with quelling it. The COVID-19 infodemic did not come as a surprise .In 2018,Heidi Larson predicted that the impact of the next major outbreak would be magnified by emotional contagion that would be digitally enabled.[13]They provide a framework for the prioritization and coordination of essential ,policy –relevant psychological ,psychiatry, clinical medicine ,behavioral and social sciences and neuroscience to ensure that any investment is efficiently targeted to crucial mental health that will need to work together in a multidiscipline fashion together with lived experience of mental health issues on COVID-19 to address these research priorities. They surveyed the public and people with lived experience of mental ill-health. The general survey done by IPSOS MORI, revealed widespread concerns about the effect of social distancing on wellbeing ,increasing anxiety ,depression ,stress and other negative feelings and concern about the including financial difficulties. [15]In this paper they have focused on impact of COVID-19 on the mental health of children and adolescents. Gender was an essential factor in the analysis of the general mental health state during the pandemic .Family and Community bonds and relationships also plays an essential role in mental health outcomes during the pandemic .For many individuals, the presence of a supportive family can be protective against mental distress. Relationships with the community ,prosocial action and sense of social responsibility also matters when it deals with mental health.[16]They have undertaken a descriptive analysis of the quantitative data. They have used stacked bar charts to depict the distribution of different infodemic categories by date and counting. They have constructed global map to examine the spatial distribution of the total count of the rumours , stigma, and conspiracy theories reported. They have performed different steps including study settings and data collection, study definition in which the definitions of rumors, stigma and conspiracy, data extraction and consistency and analysis.[17]This systematic review examined the psychological status of the general public during the COVID-19 pandemic and stressed the associated risk factors. They have also put the efforts to reduce the symptoms of mental disorders.[20]They analyzed recognized three different sub-types of misinformation that reconfigured existing information. The most common form of misinformation misleading content (29%)contained some true information, but the details were reformulated, selected and re-contextualinzed in ways that made then false or misleading .While describing the landscape of COVID-19 misinformation as an "Infodemic" captures the scale, their analysis suggests it risks mischaracterizing the nature of the problems they faced. They introduced the term misinformation moves top-down and bottom – up.[24]In this paper author designed the Copenhagen Corona related mental health(CCMH) questionnaire to focus on mental health indicators ,worries and behaviors related to other COVID-19 crisis. The questionnaires includes sociodemographics measure (i.e. age, sex, education, postal code and occupation).COVID-19 symptoms ,diagnosis, hospitalization, chronic physical and mental disorders ,perceived several social isolation ,common metal disorders crisis related mental health indicators(i.e. anxiety, lonliness, hopelessness, depression, and physical stress symptoms),quality of life, quality of sleepCOVID-19 related precautions and worries and source of information.[25]In this paper, they have categorized fake news in six different types according to their characteristics .that is malicious fake news, neutral fake news, satire news, dis-information, mis-information, rumours. [30]Data analysis is concerned with making the raw data acquired amenable to use in decision-making as well as domain –specific usage. Data Analysis involves exploring ,transforming and modeling data with the goal of high lightening, relevant data, synthesizing and extracting useful hidden information with high potential from a business point of view.

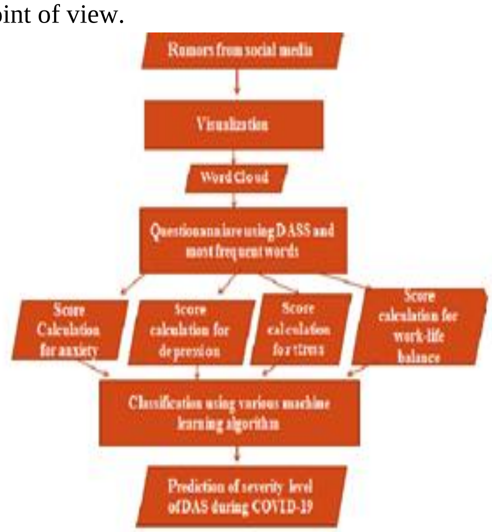

Figure-1. An architecture for the prediction of severity levels

III. Architecture

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.
Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.
1.Data Collection-Data collection is the initial step in which data will be collected from the authenticated data sources which consist of relevant attributes. The data is the mis-information related to pandemic which are collected from Infodemic through the social media platform. Datasets are downloaded from the standard sites on "rumors related to COVID-19" spread through the social media platforms. These datasets are in textual format and stored in .csv file format. These datasets contains more than 10K rumors related to COVID-19.

B.Data pre-processing phase is divided into two phases
B.1 Visualization of data during data collection
Initially, the format of data is textual and as the initial visualization techniques of data analysis we have to generate the most frequent word occurs in the textual data .The output will be the most frequent word occurs. Those occur words are the evidence of the rumors that are likely to be spread ,mostly during COVID-19 .

Figure-2 Word cloud of rumors on Social Media Platforms

B.2 Data Preparation

In this phase, the data has been prepared with the DASS-42 scale and the most frequent word that has been occurred during the visualization. The output of the phase are the questionnaires that are likely to be asked to the target population. As people are not much very comfortable to share their views regarding their mental health, publicly, so the best option for collecting their personal data is the online questionnaire. The online questionnaire will be conducted among the people those are of different professional genres.

The questionnaire contains the demographics variables (age, gender, working/nonworking, salaried/self, employed,jobtype,covid-19 victim, suffering from any disease, Income(range),belongs to any happiness group and the responses based on DASS42 will scaled between 1 to 4 and are coded as-

1) Yes, applicable to me, completely every time.
2) Yes, applicable to me, partially.
3) Yes, applicable to me, completely, sometime.
4) No, did not applicable to me.

Once the final score has been calculated, the score will be labelled, according to the affected levels, i.e. Not affected, slightly affected, mostly affected.

3) 3.1 Score calculation for anxiety

According to the questionnaire those question that has been lie under the category of anxiety the recorded responses has been labelled. For eg. If Q-1, Q-5, Q-7, Q-10 are the question based on anxiety and the score calculated are 10,25,6,14 then the respective labels are slightly affected, mostly affected, not affected and slightly affected.

3.2 Score calculation for depression

According to the questionnaire those question that has been lie under the category of depression the recorded responses has been labelled. For e.g. If Q-2,Q-4, Q-6, Q-9 are the question based on depression and the score calculated are 9,20,4,13 then the respective labels are slightly affected, mostly affected ,not affected and slightly affected.

3.3 Score calculation for stress

According to the questionnaire those question that has been lie under the category of stress, the recorded responses has been labelled. For e.g. If Q-3,Q-8, Q-11, Q-12 are the question based on stress and the score calculated are 22,5,10,6 then the respective labels are mostly affected, ,not affected, slightly affected and not affected.

3.4 Score calculation for work-life balance

According to the questionnaire those question that has been lie under the category of work-life balance; the recorded responses has been labelled. For e.g. If Q-13,Q-14, Q-15, Q-16 are the question based on stress and the score calculated are 2,13,17,19 then the respective labels are not affected , ,slightly affected, mostly affected and mostly affected.

4.Classification

Now, the classification has been carried out to predict the level of affect using the various machine learning classifiers.

The classification algorithms applied to the datasets. The dataset is used for the training and the testing and classify by using mining tools by using 5 major classification algorithm.

4.1 Support Vector Machine

As we are using multi-labelled classification techniques, The proposed techniques for multi the kernel function should have different forms and parameters. Multi-kernel SVM is a powerful machine learning tool which allows automatically to obtain suitable combinations of several kernels over several features (therefore selecting the features as well). Multi-kernel SVM is introduced to learn with multi-modality attributes or adaptively select the optimal kernel if multiple candidates exist At present, the selection of the kernel function and its parameters is based on experience, with a certain randomness. How to select the appropriate kernel function according to the actual data model in the SVM algorithm is very important problem. In this paper, we put forward multi-kernel SVM methods to adaptively select the optimal kernel for different features to get the best experimental results.[1]

4.2 K-nearest neighbour

This algorithm finds similarity between predefined classes and the classes to be classified using Euclidean distance. Another algorithm used in this category is K-star, uses similarity measure as entropy distance. The detailed description of this method is available in [18].

4.3Neural network The neural network classifier works on the principle of error correction learning. The networks learn from the training dataset and the networks evolve until an acceptable error is not found. The neural networks used in this work are multilayer perceptron (MLP) and radial basis function network (RBFN). The RBFN is more efficient because it used gaussian kernel for the separation of patterns. The description of these methods is found in [19]

4.4 Bayesian classifier

A Bayesian classifier obtains the posterior probability of each class, Ci, using Bayes rule. The naive Bayes classifier (NBC) makes the simplifying assumption that the attributes, A, are independent given the class, so the likelihood can be obtained by the product of the individual conditional probabilities of each attribute given the class. Thus, the posterior probability, P(Ci|A1, ..., An), is given by: P(Ci | A1, ..., An) = P(Ci)P(A1 | Ci) ...P(An | Ci)/P(A) (1) As we mentioned before, two important problems for the NBC are how to deal with dependent and continuous attributes.[2]

Classification is represented by using the confusion matrix. To find the accuracy of the classifier

Normally, the larger portion of the data is used for training, and a smaller fraction of the data is used for testing. We adopt fivefold cross validations for these experiments. Accordingly, we divide the collected data for each class into five equal sized folds, and use fourfolds of the labeled data for training and onefold for testing. Then we get the classification model from training data using selected classification algorithms. As shown in Table 6, we use the confusion matrix to represent the classification results. The accuracy of classifiers is measured based on precision and recall, and the F1-measure is also calculated based on the weighted harmonic means of precision and recall. Finally, we run various classification algorithms (Naive Bayes, KNN, Decision Tree and libD3C) using WEKA and compare the results with single and multiply kernel SVM method. As shown in Tables 7, 8, 9 and 10, the precision, recall, F1-measure and accuracy of Naïve Bayes, Decision Tree, KNN, libD3C, single-kernel SVM and multi-kernel SVM are listed by using different kinds of features, based on fivefold cross validation. The precision is 75.56% under the conditions of using multi-kernel SVM method based on all-features, and under the same conditions we get the F-measure (76.12%) and the recall (76.69%). We obtain better precision, recall and F-measure than other methods from these experiment results, and it is also better in all-features than part features except the recall which is lower in all-features than the microblog text features. Compared to Naïve Bayes, Decision Tree, KNN and libD3C, the precision and recall of multi-kernel SVM method based on all-features are Precision = (17) TP TP + FP Recall = (18) TP TP + FN Accuracy = (19) TP + TN TP + TN + FP + FN F1 measure = (20) 2 × Recall × Precision Recall + Precision . Fig. 4 Selection of kernel function Table 6 Confusion matrix Positive (actual) Negative (actual) (P=TP+FN) (N=TN+FP) Test outcome Positive True positive (TP) False positive (FP) Negative False negative (FN) True negative (TN) Int. J. Mach. Learn. & Cyber. 1 3 improved, and it also is found that the recall improves much more than the improvement of the precision. Meanwhile the multi-kernel SVM with all-features can achieve better accuracy of 83.46% than the other methods, and also get better accuracy than part features. We also find that the accuracy in Naïve Bayes, KNN, Decision tree and libD3C methods with user profile and behavior features is better than that in microblog text features. But the accuracy in single-kernel SVM and multi-kernel SVM methods with microblog text features is better than that in user profile and behavior features. This shows that SVM method can achieve better performance for text classification, and multi-kernel SVM method has better classification performance for multi-source and heterogeneous data in comparison with single-kernel SVM method. As shown in Table 11, by comparing with other methods, including Naive Bayes, Decision Trees, KNN, libD3C and single-kernel SVM, it is found that the error rate of Table 7 Precision of Naïve Bayes, KNN, decision tree, libD3C, single-kernel SVM and multi-kernel SVM using different features Features Naïve Bayes (%) KNN (%) Decision Tree (%) LibD3C (%) Single-kernel SVM (%) Multi-kernal

REFERENCE

[1] The COVID-19 Social Media Infodemic,Malteo Cinelli et al
[2] A first look at COVID-19 information and misinformation sharing on Twitter by Lisa Singha et al
[3] An Exploratory Study of COVID-19 Misinformation on Twitter,Gautam Kishore Shahi et al
[4] Information about the COVID-19 pandemic – A thematic analysis of different ways of perceiving true and untrue information,Stinne Glasdam et al.
[5] Managing News Overload (MNO): The COVID-19 Infodemic,Sameera Tahira Ahmed.
[6] Framework for Managing the COVID-19 Infodemic: Methods and Results of an Online, Crowdsourced WHO Technical Consultation Viroj Tangcharoensathien et al
[7] Social Media in the Times of COVID-19 Ashish Goel et al
[8] Data-Driven Methods to Monitor, Model, Forecast and Control Covid-19 Pandemic: Leveraging Data Science,Epidemiology and Control Theory,Teodoro Alamo et al
[9] INFORMATION DISORDER: Toward an interdisciplinary framework for research and policy making,Claire Wardle
[10] The spread of true and false news online, Soroush Visoughi et al
[11] Leveraging media and health communication strategies to overcome the COVID-19 infodemic,
[12] How is COVID-19 pandemic impacting mental health of children and adolescents?
[13] COVID-19–Related Infodemic and Its Impact on Public Health: A Global Social Media Analysis
[14] Impact of COVID-19 pandemic on mental health in the general population: A systematic review
[15] Types, Sources, and Claims of COVID-19 Misinformation
[16] Standing together – at a distance': Documenting changes in mental-health indicators in Denmark during the COVID-19 pandemic
[17] New Horizons for a Data-Driven Economy
[18] Viral modernity? epidemics, infodemics, and the 'bioinformational' paradigm Michael A. Peters at al.
[19] Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment, MOHAMMAD (BEHDAD) JAMSHIDI et al.
[20] Social Media and Coronavirus: Paranoid-Schizoid Technology and Pandemic? Jacob Johanssen1 et al.
[21] The spread of medical fake news in social media – the pilot quantitative study, Przemyslaw M. Waszak et al.
[22] Leveraging volunteer fact checking to identify misinformation about COVID-19 in social media, Hyunuk Kim (1), Dylan Walker (2) Affiliations: (1,2)
[23] Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention Gordon Pennycook et al.
[24] Why do People Share Misinformation during the COVID-19 Pandemic? Samuli Laato1 et al.
[25] Susceptibility to misinformation about COVID-19 around the world Jon Roozenbeek et al.
[26] Social Media can have an impact on how we manage and investigate the COVID-19 pandemic Carlos Cuello-Garcia et al.
[27] Fake news and COVID- 19: modelling the predictors of fake news sharing among social media users Oberiri Destiny Apuke et al.
[28] The Parallel Pandemic: Medical Misinformation and COVID-19 Primum non nocere Jennifer S. Love et al.
[29] Data-Driven Modelling: Concepts, Approaches and Experiences D. Solomatine,et al
[30] Data-driven modeling and learning in science and engineering Francisco J. Montáns et al.
[31] Innovation: A data-driven approach Andrew Kusiak et al.
[32] DATA SCIENCE AND ITS RELATIONSHIP TO BIG DATA AND DATA-DRIVEN DECISION MAKING Foster Provost et al.

[33] Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis Fred Mannering et al.
[34] Big data analytics in healthcare: promise and potential. Wullianallur Raghupathi et al.
[35] Initial psychological impact of COVID-19 and its correlates in Indian Community: An online (FEEL-COVID) survey Mohit Varshney et al.
[36] The Impact of Event Scale-Revised: Psychometric properties in a sample of motor vehicle accident survivors, J. Gayle Beck et al.
[37] The Impact of Event Scale-Revised: Psychometric properties in a sample of motor vehicle accident survivors J. Gayle Beck et al.
[38] COVID-19: Psychological flexibility, coping, mental health, and wellbeing in the UK during the pandemic David L. Dawson et al