International Journal of E-Health and Medical Communications
Volume 13 • Issue 2 • July-December 2022

Semantic Pattern Detection in COVID-19 Using Contextual Clustering and Intelligent Topic Modeling

Pooja Kherwa, Maharaja Surajmal Institute of Technology, Delhi, India
Poonam Bansal, Maharaja Surajmal Institute of Technology, Delhi, India

ABSTRACT

The COVID-19 pandemic is the deadliest outbreak in our living memory. So, it is the need of hour to prepare the world with strategies to prevent and control the impact of the pandemic. In this paper, a novel semantic pattern detection approach in the COVID-19 literature using contextual clustering and intelligent topic modeling is presented. For contextual clustering, three level weights at term level, document level, and corpus level are used with latent semantic analysis. For intelligent topic modeling, semantic collocations using pointwise mutual information (PMI), and log frequency biased mutual dependency (LBMD) are selected, and latent dirichlet allocation is applied. Contextual clustering with latent semantic analysis presents semantic spaces with high correlation in terms at corpus level. Through intelligent topic modeling, topics are improved in the form of lower perplexity and highly coherent. This research helps in finding the knowledge gap in the area of COVID-19 research and offered direction for future research.

KEYWORDS

Latent Dirichlet Allocation, Latent Semantic Analysis, Log Frequency Biased Mutual Dependency, Mutual Information, Point Wise, Vector Space Model

INTRODUCTION

The Coronavirus family comprises of a wide range of animal and human viruses. Coronaviruses are positive-sense RNA viruses and are classiðed into four genera: Alpha, Beta, Gamma, and Delta-coronaviruses. (Weiss & Leibowitz.,2011; Burrell et al., 2016). Alpha coronaviruses and beta-coronaviruses are found exclusively in mammals, whereas gamma coronaviruses and delta-coronaviruses primarily infect birds. Prior to 2003, members of this family were believed to cause only mild respiratory illness in humans.

The 2003 epidemic of SARS-Cov prompted an intensive research for novel coronaviruses, resulting in the detection of a number. of novel coronaviruses in humans, domestic animals and wildlife. This research finds the greatest discovery, which suggest that bat and avian species are the natural reservoirs of the viruses (Guo,2020). Recent studies also discover that these coronaviruses are the result of recent cross species transmission events.

DOI: 10.4018/IJEHMC.20220701.oa7

The emerging of novel corona virus(2019-nCov) has awakened the echoes of SARS-Cov from nearly two decades ago (Gralinski & Menachery 2020). This zoonotic human coronavirus of the century emerged in Dec 2019, with a cluster of patients with connection to Huanan south China sea food market in Wuhan, Hubei Province China. Similar to severe acute respiratory syndrome (SAR-Cov) and Middle east respiratory Syndrome coronavirus (MERS-Cov) infections patients exhibited symptoms of viral, pneumonia including fever difficulty in breathes and bilateral lung infiltration in the most severe cases (Wuhan Municipal Health Commision,2020).

Since its emergence in China in Dec 19, the coronavirus is spreading very fast in the entire world. Till 8th June globally there have been 6,881,352 confirmed cases of COVID-19, including 399,895 deaths, reported to WHO. According to country wise detail data from WHO dashboard on 8th June 2020, United states of America has the highest number of confirmed cases of 1915712, And at second largest confirmed cases in Brazil with 672846, and then Russia is at third position with 467673 and United Kingdom with 284,872 confirmed cases. Till 8th June India has total 265740 confirmed corona cases, out of which 129358 are active cases and 128894 recovered successfully, and unfortunately 7473 deceased [https://www.covid19india.org/]

India as the 2nd largest most populated country of world after china, where the first Covid-19 case emerged in Kerala on Jan 30,2020, which originally originated from China. Till 20 March, India observed around 223 confirmed cases and out of 4 lost their lives due to this pandemic. Indian Government has taken all the necessary step to tackle the pandemic in our country.

Till today the Covid -19, pandemic shows no sign of abating, as vaccine is yet to found. Although all the countries are trying to control with lockdown and local and global social distancing. Even in some countries situation is under control, Government started unlocking the country in phases with necessary precautions.

The researchers in different part of worlds are trying their best in different research labs and individually in different fields of medicine, bioinformatics, virology, technology, Data analytics, artificial intelligence to help the humanity to tackle this horrible epidemic with minimum loss.

Data scientist and analytics with advanced machine learning, deep learning algorithms try to predict the number of infected people in the future, also try to predict number of susceptible populations, so that government can take necessary action like implementation of lockdown, building necessary healthcare infrastructure.

In this paper, our approach towards Covid-19 pandemic is using distributional semantics, here the emphasis is to present semantic pattern in available literature of Covid-19, through contextual hierarchical clustering and intelligent topic modeling. For contextual hierarchical clustering implementation latent semantic analysis with novel three level weights at term level, document level and corpus level are used. We choose two three level semantic space, ATC – Augmented weighting at term level, log term frequency at document level and Cosine normalization at corpus level and, NPC-Neutral at term level, probabilistic weighting at document level and Cosine normalization at corpus level.

Intelligent topic modeling is implemented using semantic collocations selection using point wise mutual information (PMI) and log frequency biased mutual dependency (LBMD) and then latent dirichlet allocation is applied. To show the effectiveness of our proposed methodology, both the approaches are compared with neutral weights at three level in contextual hierarchical clustering and traditional topic modeling algorithm latent dirichlet allocation.

The paper begins with data collection understanding, followed by 4 stages of analysis

1. Keyword trend analysis.
2. Contextual Hierarchical clustering in three semantic spaces
3. Cosine Similarity score analysis of term pair in three semantic spaces
4. Topic Modeling of Dataset using Intelligent Latent Dirichlet Allocation.