CHAPTER TWELVE

Lexical modeling and weighted matrices for analyses of COVID-19 outbreak

V. Kakulapati[a], Sheri Mahender Reddy[b], and Nitesh Kumar[c]
[a]Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India
[b]Otto-Friedrich University of Bamberg, IsoSySc, Bamberg, Germany
[c]Centurion University of Technology and Management, R. Sitapur, Odisha, India

1. Introduction

COVID-19 is transmitting to all people, even newborns who are at risk. People with severe health debilities such as hypertension, diabetes, chronic kidney disease, respiratory problems, and cardiovascular illness are at higher risk of contracting the infection and developing severe complications. Medical professionals and WHO recommend that the population at any age group the appropriate measures against the virus, such as hand washing, protecting the nose mouth, proper exercise hygiene, and decontaminating conditions. The significant probability of the transmissible virus for senior persons with disabilities with persistent disorders is considerable. The aging of human immunity steadily decreases its ability to respond to the disease owing to other health problems.

Pulmonary infection caused by a new Coronavirus, the 2019 Coronavirus, has spread globally across 222 countries. Since April 26, 2021, 148,145,493 COVID-19 recognized cases and more than 31,271,098 mortality worldwide, based on the current World Health Organization (World Meters Statistics).

In order to interpret immunological test cases of infections, computational formulating of diagnostic biomarkers is immensely useful and allows us to make efforts to prevent disease spread by reducing the threats of global health. In addition, computational analysis and biotechnological investigation mostly need to cooperate and significantly eliminate the outbreak of COVID-19. The prediction of the outbreak and termination time of the disease is completed by numerical techniques. It allows us to decide correctly on the actions needed to limit communicable diseases. Unfortunately,

Lessons from COVID-19
https://doi.org/10.1016/B978-0-323-99878-9.00005-4

COVID-19 spread rapidly around the world in 2019. Then it is a serious predicament since the prevalence of SARS-COV-2 is high (World Health Organization, 2019). Therefore, infectious statistical data are essential to increase awareness and inform engagement (Rivers et al., 2019).

In diagnosing the disease, determining its epidemiological, patient treatment, and reducing propagation, the screening test of the COVID-19 is essential. Standardized policies and methodologies harmonize existing medical tests regarding efficient diagnostic methods in the worldwide battle against the disease outbreak. Furthermore, the specification of even more severe acute respiratory variants is entrusting scholars, practitioners, and biologists to enhance the cluster-based precision and accuracy of antibodies and diagnostic testing. Easy, scalable, fast, and cost-effective diagnostic procedures for COVID-19 can substantially assist in creating high sensitivities, which provide efficient screening capabilities in the outbreak on-demand. It is a fundamental scientific reference for precise estimation of the diagnosis difficulties of COVID-19 and enhanced illness fight techniques (Kevadiya et al., 2021).

Topic modeling tools are beneficial when a community combines to realize the contamination activities and calculate its removal or survival conditions. Scientific research, government, and civilians are now very anxious about the high transmission rate of the infection and the significant mortality rate. Coronavirus is an infectious ailment first recognized in December 2019 by a Coronavirus in Wuhan. When a patient coughs or sneezes excessively, the COVID-19 virus spreads through droplets. These droplets rapidly spread in the air and also settle on the surfaces.

Proposing a new framework and system that uses ML approaches to retrieve significant and relevant characteristics from the COVID-19, including nonnegative matrix factoring for NLP. Linguistic approaches have been influenced by computational linguistics for an extended period. Over 100 years, probabilistic reasoning influenced linguistics (Frege, 1967), and a prominent origin in the computational modeling of languages is conceptual grammar, and among the most commonly recognized language models (Levelt, 2008).

The SVM is a nonlinear method of machine learning for the classification of entities using analysis methods. SVM intends to find the best hyperplane which identifies all input variables with one category from the observations in another (Cortes and Vapnik, 1995).

In order to achieve optimum precision, the SVM method to forecast an occurrence of COVID-19 is developing. COVID-19 estimation SVM classification analyses a numerous cases dataset that enables particular capabilities for every collection. The objective was to establish an SVM-based model to predict epidemics that could analyze the COVID-19 virus and provide a testing process before this analysis. The SVM approach is used to produce the central perspective of a small training set, particularly in the context of the outbreak assessment. The SVM model is also being used to ensure the highest precision to anticipate a new Coronavirus pandemic.

The AUC-ROC curve is a classification problem, which requires continuous improvement at different thresholds. ROC is the probabilistic progression, while AUC is the entropy measure of the extent or measurement. It indicates whether the classifier can differentiate among classes. The increased the AUC, the better the classifier. The higher the AUC, the better the model determines among patients and no disease.

The semantic technique builds the concept of sentiment on the analysis of individual words or expressions; emotional lexicons are typically using: they provide an accessible emotional vocabulary from the dictionary, estimate sentiment weights, and use a combined weight function.

LDA topic modeling is a popular sampling technique for allocating enormous datasets into semantically relevant subgroups. Files are arbitrarily chosen variations of subjects in the dataset that constitute the reasoning between topics special for a given subject. The President of the United States, for instance, has published a news story about healthcare. The issues in the news would be suitable for the president of the US, health insurance, and political beliefs. However, they convey LDA medicine in language that might exist in more than one topic and maybe joint into issues.

The model addresses the issue by analyzing a document on topics and identifying that concept is closer to the content laterally on all subject variations. It helps the keywords to understand which documents are the topics. PLSA is equivalent to LDA, excluding the subject; allocation should be SDP. SDPs detect if papers cover a few subjects; just a few words are used often in these areas. The consequences are a word pattern and then the identical problem formalities. The findings contained the LDA generalization of the PLSA model (Kakulapati and Reddy, 2020).

Statistical models used to anticipate infections can discuss the severity and preventability of the diagnosis. It is utilizing to guide and assist pick the proper health decisions. The proposed method has important implications. First, the classifier precisely analyzes disease patterns and impacts to enhance diagnosing and assessing physicians' therapeutic urgencies. Second, the process utilized to estimate the many factors disturbing the transmission of the ailment is prognostic and determines the efficacy.

This work intends to build and validate a precisely predicted machine learning model

- The model is predicting the Coronavirus patients and older people with underlying diseases by using SVM with AUC.
- The model proposed is a framework of recognition based on the lexicon. It integrates awareness with the intensity of emotional lexicons to accurately recognize the substance of feelings in words and accumulate exact feeling knowledge.
- The proposed model understands the interactions less noisy than existing treatment approaches, using classification techniques. Therefore, the words in the SVM model are summarizing and analyzed.
- The model organizes the latest advanced techniques for the assessment of words in datasets. The impact of integrating lexicons of perception and LDA with word cloud was visualizing.

The clinical evaluation facilitates the diagnosis of COVID-19. Severe acute respiratory syndrome Coronavirus 2 viruses can be anticipating more correctly based on the kind and mix of symptoms. It can allow for a fast diagnostic evaluation of the illness in a high prevalence environment. Cough, myalgia, smell/goodness loss, and fiver over 37.5°C have a high positive preview rate for infection with SARS-CoV-2 (Mardian et al., 2021).

The remaining sections of this chapter are explaining effectively. Section 2 discusses past studies addressing the lexical interpretation techniques and the LDA modeling of linguistics. Section 3 elaborates on some diagnostic measures of COVID-19. Then discuss detailed methodology and its training processes, including the system architecture in Section 4. Section 5 offers experimental findings for several language pairings utilizing the proposed lexical interpretation model of learning applications. Section 6 discusses our model assessment methods. Section 7 concludes. Finally, Section 8 explains our future updations.

2. Related work

Developing a suitable computerized screening test for medical professionals might facilitate immediate diagnosis assessment, for example, formalized legitimate examination of relevant signs for COVID-19. To reduce further transmission at the initial stage of the virus, this is essential to have an early warning system considering a diagnosis of different numerous medical factors (Van Loon et al., 2021).

Blood samples are taken from patients to identify the Coronavirus with the RT-PCR method (Chan et al., 2020). Patients tested the impact of (RdRp)/Hel) and found that COVID-19 identification (RdRp)/Hel) inpatient cases also do not interfere with other bacterial pathogens and biological Coronaviruses. The author stated the potential to affect public welfare via strengthening the COVID-19 lab diagnosis and treatment in this thorough diagnostic screening.

Predict the probability of patients with underlying issues having influenced Coronavirus applying an ML algorithm to estimate the disease prevalence and serious complications. Simulation of outbreaks with this forecasting technique includes a massive range of possible conditions. Forecasting can elevate the interpretation of demographic changes and unpredictable age structure and recommendations for pandemic avoidance. Although COVID-19 causes minor infections among the majority of the patients, it does require special attention for the mortality risk of patients with preexisting diseases (Kakulapati et al., 2020).

The RT-PCR recommended validating diagnoses for COVID-19, which can supplement a diagnosing medical analysis and antibodies and antibodies tests. Although antimicrobial testing mainly suggests antigen detection because of the slow onset of symptoms, antigenic testing can quickly identify severely contagious patients in the illness trajectory that can considerably limit spread (Mardian et al., 2021).

Linguistic approaches, directive approaches, ML techniques, hybrid techniques, and multilabel techniques are inherent in different human classifiers. In Al-A'abed and Al-Ayyoub (2016), a lexicon-based Twitter and Facebook emotional language evaluation method has been establishing. They revealed that 89.7% of the lexicon-based strategy is operative and attain an accurateness of 90%. In another investigation, sentimental reagents were evaluating in the Coronavirus epidemic by analyzing tweets (Kaila and Prasad, 2020). The eight sentiments were studying together with a subsample of tweets. The consequences enable the identification of optimistic and pessimistic beliefs that are almost equal. The patients' anxiety was the leading emotion, followed by governments' optimism, which influenced the tweets.

Twitter messages were evaluated in linguistics (Abd-Alrazaq et al., 2020) during February 2, 2020, and March 15, 2020, and assessed the tweets were retrieved in single (uni-gram) or dual (bi-gram) keyword occurrences. Twelve topics have been recognized and classified into four major topics: the origin of the epidemic, potential effects on individuals, regions, the financial system, and measures of minimizing infection risk. Emotion recognition, identity verification, sentiment classification, etc. are different applications for text summarization. In addition, opinions of Indian government policies using a Lexicon Vocabulary were investigating (Verma et al., 2019).

A methodology is known as a dynamic analysis of NLP. It aims to recognize and remove the polarization of feelings from information by establishing whether the test is positive, negative, or neutral (García-Díaz et al., 2020). Two sentiment analyzing methodologies exist: machine learning and lexical-based analysis. Machine training employs algorithms to detect feelings, but the approach to lexicons is positive and negative (Drus and Khalid, 2019).

Many investigations have been carried out in sentiment classification, utilizing a computational intelligence approach by (Zhang et al., 2018), described in (Jelodar et al., 2020), employing a computational intelligence model to classify emotions. The NLP is used to analyze topics to determine the underlying issues associated with COVID-19 on social sites. Objects are organizing by utilizing the LSTM RNN model.

The n-gram-derived method (Acebal et al., 2006) provides a valid approximation of the Integrated Probability of origin and destination text from the maximum likelihood of a specific n-gram combination. The basics of a conventional linguistic string of words or n-gram models are mathematical concepts such as topic modeling and thinning methods. However, mathematical models have increasingly been enough to learn simple models in which nonlinear linguistic interactions and contextual word classification occur (Mikolov et al., 2013). The n-gram interpretation method (Hai Son et al., 2012) describes the model as a consequence of a bidirectional phrase paired sequence with a perceptron, the probabilities of a word sequence. Rebuild the bilateral technique (Devlin et al., 2014) as an alternative to the KNN rescoring module in the n-gram decoder instead of using the KNN scorer. The maximum likelihood (Schwenk, 2012) is then adapted to the word sequence in a source sequence.

In the epidemic COVID-19 in China, a statistical model was developed by Yang et al. (2020). The pandemic's intensity and frequency and the maximum epidermal aspect are predicted by a simplified deterministic concept of susceptible eroded infectious recovered. This is a significant consequence for investigating the propagation and spreading of COVID-19 utilizing computational mathematics methods.

Modeling topics analyzed massive datasets of geotagged tweets (Dahal et al., 2019), including global warming using topic modelization and text analytics. In retrieving the topics in the text, LDA uses feature selection and VAD (Valence Aware Dictionary and Sentiment Reasoner) to assess fundamental emotional reactions. Emotion recognition illustrated that if users respond to government or weather patterns incidents, most conversation is unfavorable. In contrast, topic modeling showed that numerous such topics of climate change discussions are broad but more significant.

The LDA approach used to analyze new COVID-19 epidemics using facts acquired indicates that people are scared of new Coronaviruses' worldwide epidemics (Kaila and Prasad, 2020). Muthusami et al. (2020) established the COVID-19 epidemic to evaluate emotions utilizing various ML algorithms. Bakshi et al. (2016) used the NLTK library to clean and text to analyze tweets for categorization of public opinion. Several organizations expressed their opinions on Twitter.

Analytical methods have for many decades generated probabilistic insights and perspectives for epidemic decision-making and policy incorporation in the epidemic (Wu et al., 2020a). Several experiments conducted for COVID-19 specifically are performed. For example, COVID-19's description model is produced by transmission changing aspects in China and anticipates global and regional disease transmission data reported in China. Read et al. (2020) discussed the initial outbreaks' effective atomic rate, utilizing a hypothesis of regular incremental Poisson's latency for information transformation.

The sentiment evaluates a particular segment of the text's emotions, feelings, opinions, and attitudes (Medhat et al., 2014). A fundamental objective for emotional investigation is to examine if a term in a specific aspect is favorable or unfavorable (Ravi and Ravi, 2015).

Researchers use consistent weights (Nurhayati et al., 2012), binary weights (Mukhaiyar and Pasaribu, 2012), distance-based weights. This methodology of identifying weight persists qualitatively. The specified weight matrix infers better correlations among areas. This matrix is a hybrid of the reasonable inverse distance weight (IDW) matrix with the patient ratios among provinces, and the IDW matrix changed.

3. About COVID-19 diagnostic methods

In diverse hosts like animals and people, Coronavirus causes moderate-to-severe infectious respiratory illness, transfers, and survives in various hosts. The virus is detected in patients with high temperatures, intense muscle aches, lung congestion, sputum production, respiratory distress, dyspnea, hemoptysis, and diarrhea. In addition, anosmia and ageusia have been documented.

It is recommended to take a throat swab to assess and evaluate the infection in the initial evaluation of COVID-19. In order to prevent tongue injury, the rear of the pharynx should be swabbed. On the contrary, pulmonary lowering samples are recommended for use in saliva, low respiration aspirates, and bronchial washing (Pan et al., 2020). Thus, to increase the sensitivity of the testing and save access to a good combination of NP and OP swabs and repetitive sampling.Consequently, between samples, the accuracy of the NP swab may not be consistent. Physicians may also be susceptible to infection in the procedure of testing due to aerosols production (Guo et al., 2020).

Additionally, the requirement to separate COVID-19 from seasonal influenza instances may delay emergencies, facilities' slow care and treatment for persistent with coronavirus, and constitute a significant source of apprehension (Liu and Pan, 2020). Such problems affect the scientific method of whether diagnostic procedures are best for combating the impending growth of COVID-19 infections effectively and identifying difference between COVID-19 illnesses and seasonal influenza.

3.1 Different diagnostic tests

Diagnostic procedures can show that you may have an infectious disease with COVID-19 and respond appropriately to quarantine or detach yourself. The diagnoses are types that can be determining whether you have an autoimmune disorder of COVID-19. Diagnosis specimens are often collected by spitting into tubes with a nasal or mouth swab or saliva taken.

In antibody diagnostics, the SARS-CoV-2 virus that further stimulates COVID-19 is used to detect antibodies in the human immune system. However, the analysis of persistent COVID-19 contagion is not completed by antibody analysis. Once infected, immunoglobulins can be taken several days or even weeks and can last many weeks or more in human blood after recovering. Medical professionals collect blood samples from the patients for the detection of antibodies.

Worldwide, government agencies have authorized numerous diagnostic procedures; nevertheless, the right policies are still unclear, dependent on the health status or the objectives of the research of the patient (Sanyaolu et al., 2020). For example, in Europe, the CE-IVD and FDA authorized or utilized for data analysis 365 different commercialized tools. Some are immunological investigations, polymerase chain reaction procedures, NGS-based procedures, and two marketed instruments depending on other technologies (European Commission, 2020). And, therefore, it is transparent that it can be challenging to select accurate therapeutic testing. The test result may be positive or negative for COVID-19.

3.1.1 COVID-19 positive

In other words, people are currently suffering from an infectious illness with the COVID-19 virus. Taking immediate action to protect themselves and prevent the spread of infection to others. People have to isolate themselves until the symptoms disappear and they have to have an average temperature for 24 h and at least 10 days after the first appearance of signs or are proven positive, whichever is later. The patient recommends staying in social exclusion extended if they have symptoms associated with COVID-19 or have a health status that affects their illness control. If the result is positive, and no signs or symptoms have developed, isolation 1 week after the testing.

3.1.2 COVID-19 negative

The COVID-19 virus probably didn't contaminate. Depending on the test sample's efficiency and effectiveness, a fake negative test result may occur. If the patient is already tested negative and may get contaminated in perspective, it is essential to recognize isolation rules, masks, and washing hands to prevent possible spreading. If the patient continues to have signs, the doctor may prescribe repetitive tests.

3.1.3 Polymerase chain reaction test

This COVID-19 test is also known as a molecular test and examines the virus's genes to use a scientific technique known as PCR. A fluids spectrum is collected with insertion into the patient nose of a nasopharyngeal swab and fluid from the patient nasal or a mediate turbine swab to test for the presence. An oropharyngeal swab is deposited in a test tube, saliva in the mouth, or split into a tube in rare instances. If investigated on the spot, or some few consecutive days—or more in regions with testing delays—findings may be received in minutes when transmitted to individuals impacted. PCR testing is quite precise if appropriately performed by a healthcare provider, although specific instances may miss a quick test.

3.1.4 Antigen test

In the COVID-19 test, various proteins in the virus are detected. Specific antigen testing can deliver findings in minutes to use a long nose swab for a fluid sample. These are sent for analysis to a laboratory. However, if the guidelines are executed precisely, a positive antigen testing procedure is recognized accurately, and there is a rise in the possibility of wrongly negated results. That is, the infection is contaminated, but the outcome is negative. Therefore, a PCR test to validate negative antigen testing may be recommended (Mayo Clinic).

It enables precise control of synthesized antibodies, a profound understanding of antibodies generated by various tissue contexts, and may cause false negatives if the development of gene products is limited.

3.1.5 RT-PCR-based molecular tests

The molecular RT-PCR procedures are the golden standards utilized worldwide to determine COVID-19 infection in a confirmed manner. Investigators had already completely sequenced the SARS-CoV-2 genome (Wu et al., 2020b), initiated the structure of biological primers and samples precise to the SARS-CoV-2 RNA molecules to analyze COVID-19 infectious disease and some other similar symptoms, including regular influenza or contagious bacterial infections. The entire SARS-CoV-2 genomic sequencing has a frequency of 29,903 bp. It contains the specific functionality: A 50 base Poly-A cap, 1/ab opening with genome polymerization RNA, spike proteins, membrane receptor, molecule and nucleocapsid protein, and 35-bp polyA tail synthesis. Time required for the test; qualified professionals and adequate infrastructure for data interpretation.

3.1.6 Chest computed tomography

The scan results are noninvasive and comprise several X-ray observations of the chest to collect cross-sectional images (Whiting et al., 2015). Primary care physicians investigate the images to identify abnormalities that can result in diagnosis. The radiological features of the Coronavirus vary from stage to stage and depend on the severity of the infection. For instance, in the initial phases of the disease (0–2 days), more prevalent standard CT results with significant respiratory infiltration rise 10 days after symptoms. The general characteristics of COVID-19 consist mainly of the bilaterally symmetrical and peripheral opacity zone in the lower lobes and pulmonary congestion (Bernheim et al., 2020). Numerous retrospective investigations showed that CT scans are much more precise and had enhanced undesirable result proportions than RT-PCR, depending on those scanning features.

Loop-mediated isothermal amplification: By relying on certain SARS-CoV-2 genomes, the possibility of active infection exists. It is analytical and uses reagents that are precisely framed to construct new sequence structures for replication. It is quick and needs no extra safety equipment.

However, the limit of detection of 125 viruses/mL is extremely low. The development of the appropriate reagents might be complicated, and the reactivity of waste can intervene. The consequences can also really be a measured level of viral infection.

3.1.7 Robotic process automation

It is used to handle particular difficulties arising from the epidemic. COVID-19 analysis involves many repeated administrators, including system identification and registration of patients, accurately designated test kits, and recording test findings into national databases with millions of other entries in certain situations. In addition, a great demand for testing in many nations has resulted in shortages, although social separation provisions for employees remain valid.

These are COVID-19 diagnostic procedures, and several other ways are available, such as CRISPR, RT-LAMP, Droplet Digital PCR, Serological techniques, etc.

4. Framework

Creating a lexical framework to define how medications, diagnostics, and related information coexist in coronavirus health data. E-enhanced the accuracy of the data to analyze proportional medicines depending on the current proposed approach. Diagnostic tests may be cooperatively acceptable. To get the fundamental features and prediction models, use a lexical framework for weight matrices.

The elements of the framework (Fig. 1) are discussed below:

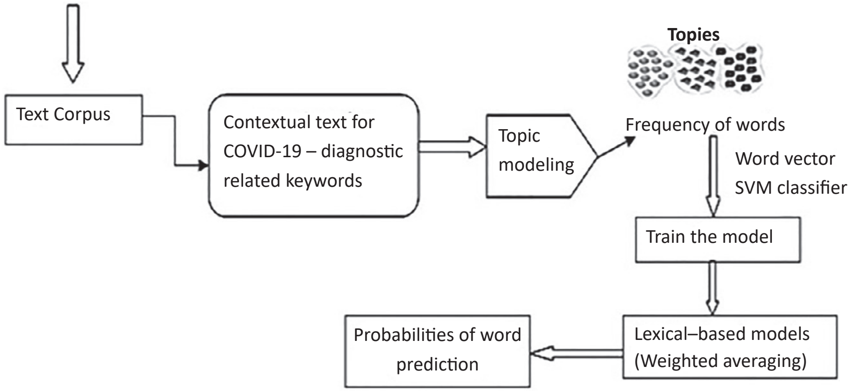
Topies
Text Corpus
Contextual text for COVID-19 – diagnostic related keywords
Topic modeling
Frequency of words
Word vector
SVM classifier
Train the model
Lexical–based models (Weighted averaging)
Probabilities of word prediction

Fig. 1 The framework of the proposed method.

4.1 SVM

It is a technique to categorize texts into distinct groups. SVM is a supervised learning classification. For the particular text, the label takes the "n" number of characteristics. It is an effective regression-classifier and significantly correlated with the dataset supporting the significant stake of efficient data (Mayo Clinic). Therefore, analyzing the model's performance becomes one of the essential priorities after data analysis. SVM is a supervised vector theory learning method. The data are classified in vectors in the area to categorize the data using this methodology. Hyperplanes are utilized for the decision-making and classification of data points by keeping the various categories from one another as far as feasible.

4.2 Lexicon intelligence

Lexicons imply accumulated perspective data that provide a foundation for conceptualization. While lexicon data are commonly assumed to play a function in analyzing linguistic (Falkum and Vicente, 2015), it is unknown in the context of every phrase. For instance, it can be set aside uniquely in the lexicon or the frequency of the interpretation affecting the lexical entries and its method—lexicons in a feature vector utilized distributed word form or incorporating. Term integrations are observed as a lexicon module that performs as a lexicon. The more identical the variations of two words are, the exact equivalent the phrase interpretations are optimized.

Possible descriptions (Boleda, 2020) are demographic semantic models of the word weight, especially LSA (Landauer and Dumais, 1997), suitable for patients using different word modeling methodologies. This is probably because (a) the observational effectiveness of the model is to describe lexicons via word embedding in a variety of tasks, and (b) that the same allow us interoperable continue providing linguistic and historical data because that can also investigate every parameter impact in isolation and permutation. Thus, in specific terms, the utilization of linguistic information is a selection of modeling.

There are various methods for creating topics based on documents. Modeling is one of the most useful. This chapter built an LDA model that can be used in multiple processes to allocate subjects. To better understand modeling topics, models are assigned to determine an issue in datasets. The topic modeling technique is important as theme models make it possible to decide on a matter in datasets.

A key issue is to provide an approach to determine similarities between crucial phrases and emotive linguistic concepts. With this goal, we computed a massive data word similarity. Especially in obtaining the linear word vectors, representation and syntax applied a derivative from LSA. Through an SVD, T1 term cooccurrences in corpus documents (Deerwester et al., 1990) are acquired utilizing the decrease in the word by document matrix.

4.3 Feature extraction

The high-dimensional data (Ahmed et al., 2017) are one of the summarization concerns. The valuable advice of morphing words into matrices (or vectors) of features is to use specific extraction process methods. It utilized two of the most influential methodologies for retrieving attributes: n-gram and TF-IDF, utilizing chronological tweets. The n-gram model is a popular method in text mining and NLP for testing and validation.

An n-gram is utilized to process correlative phrases in a prescribed period as per linguistic, statistical analyses. In this analysis, the n-gram process such as $n = 1$ to $n = 4$ was working to recognize the data context (i.e., unigram, bigram, trigram, and quat-gram). The concept TF-IDF is an effective method of evaluating the importance of a word or phrase that extracts data and utilizes NLP. The weighted text frequency is employed. In an extensive collection of documents, TF-IDF aims to identify the linguistic features inside the text. The TF-IDF methodology uses the degree of the relative frequency across the corpus of the documents, which is quite effective.

4.4 Tokenization

Preprocessing tokenization refers to the splitting of long feature vectors into tokens. These tokens may be phrases that are further dividing into small terms then subdivided into words. For example, the phrase: "hard effort pays off before the tokenization."

4.5 Normalization

The normalization process in data preparation transforms a word into a simple form for elevated linguistic consistency. It contains the transformation to the upper or lower case of all text.

4.6 Stemming

This is subsequent step in the tokenization phase. The next step is to alter parent words to reduce keyword types or categories.

4.7 Word vectors

These express words significantly more efficiently and have a significantly lower space than a smooth validity, encode vector, and contain semantic word information. These are the primary features of word vectors. Prior to proceeding into determining word vectors mystified, it is necessary to recognize the critical aspect behind the concept of word vectors (Deep Learning Demystified, 2021).

Lexical items become more frequent than different word combining.

- When a word is around, it does not always signify that it has a comparable significance.
- Words with similar meanings are regularly using when we evaluate the frequency of terms found closer together.
- Utilize a single phrase, say the nth term, only effect the nth row in the weight matrix, as all other rows are 0.
- The nth row of the weight matrix has all the learned information about the nth word, and for the nth term, this row is the trained word vector.

NMF is frequently employing for high dimension data such as summarization. Since it consistently generates sparse, significant features from a set of primary input vectors (Xu et al., 2003). It uses nonnegative weights, which are common approaches including Term Frequency – Inverse Document Frequency.

1. Matrix A = Matrix W * Matrix H
2. A = DTM for m and n words of textual data (m X n)
3. As an input, NMF takes A and factorizes it into two nonnegative matrices, W and H, with dimensions k, i.e., several topics.
4. W = In the matrix (m × k) for K-topics, the W rows weigh many of those values reveal the importance of the link between documents and subjects in the test document.
5. The topic (k × n) matrix column is H, which offers weights for phrases (columns) concerning the subjects (rows). By picking the data in a particular column and selecting the terms at the top, we may create a description (topic) of the subject.
6. The document-topic matrix is proportionally scaled: Measure the document-topic matrix so that the topic values associated with every document add up to one to evaluate and illustrate components of NMF comparable to LDA's topic proportions.
7. Measure the topic matrix correspondingly: Measure the topic matrix so that the term values for each subject are one.

5. Implementation result

Text corpus consists of the verbal information from elderly persons as well as doctors. It also includes the data from the COVID-recovered patients.

All of the information gathered is modeled using TM, and probable topics-terms are extracted. These show us the pertinent information regarding the terms frequently used around the COVID patients. Using SVM for the vectors obtained from topic modeling, we classify the vectors for precise information from the recovered patients, doctors, and others with respective probabilities.

The infection of COVID-19 can vary from mild to acute with indications such as temperature, cold, coughing, shortness of breath, exhaustion, anxiety, tasting or smell loss, chest infection, mucus, sneezing, nausea, and severe dehydration. Acute diseases, increased morbidity, and death may result.

This information may be helpful in decision-making during the spread of COVID.

Data: Coronavirus pandemic classified as a health emergency by WHO investigators, and healthcare facilities openly access the pandemic data. Two datasets have been collected and incorporated from the Kaggle open-source data repository. The dataset contains the patient information showing preexisting diseases and age.

5.1 Prepossessing

The gathered data are prepared to create the subsequent stage during this stage. This process consists of three major phases. First, remove duplicated characters, corrective text, normalization, elimination of words, and linguistic recognition in a cleaning step. A further objective of the tokenization approach is to transform the text into tokens once it generates vectors—lastly, linguistic patterns and extracting features.

During lexicon definition, the actual case documents are preprocessing to apply topic modeling to describe individual patients in a vector format such as analogous to vector documentation. The word "vector" reflects the numerous times a subject occurs in a patient file. Repetitive conditions repeat in a file; for example, chronic conditions generally signify a significantly greater intensity, leading to increased weights. The condition that only occurs occasionally in the patients' information, such as sometimes coughing or headache, is lowered and less influenced by the topic profiles.

There are five main steps in implementation: text cleaning, tokenization, stop word removal, lemmatization, and persistence. Retrieving certain features from the text data should be an input for classification algorithms in a precisely considered form. In the retrieval, numerous approaches are used. TF-IDF is applied to estimate a bag of words that separates text into words for a vector (Sethi et al., 2020). If there is one word, the approach discusses it as a unigram, the two-word as bigram, and the three-word as trigram (Chakraborty et al., 2020). The vector is extracted from this model using "unigram" approach. While executing the LDA model, several parameters need to be defined. Essential factors that should be adequately choosing include the number of topics (k) and the number of iterations that affect several times LDA executed through every text. In addition, it is also essential to develop the sample size according to the given dataset. Thus, this procedure is carried when the model executes, and K values are changed until the optimum outcome is obtained.

5.2 Evaluating metrics

The classifiers' performance evaluates with precision, sensitivity, and specificity. In each cross-validation fold, the three metrics were estimated. The correlation coefficients were gradually aggregated—assess the classification accuracy utilized for the ROC (receivers operating characteristic) analysis. Thus, a measure of classification effectiveness is the AUC (area under the curve) measured from the ROC curve.

5.2.1 Precision

It's the performance measure. Precision offers the fractional value for the total samples in a dataset of the properly predicted samples. Precision measurements correctly recognized the fraction of real positive.

5.2.2 ROC curve

The efficiency of the classification model is calculated using a measure. AUC represents the area under the ROC curve. AUC underlines how effective it is that the optimistic categories are isolated from the pessimistic.

5.2.3 N-gram

An n-gram model of language given the word sequence w,

$$P\ (wi|wi-1, wi-2, \ldots, wi-n+1)$$

where wi is the ith word of the sequence.

Quantify matrices of similarities in combinations such that associated words achieve strong effects of similarity. Then, based on the input text, employ lexical similarity. F or all word combinations, the correlation method returns a value between 0, which shows a few words are not correlated, and 1, which implies that they are pretty similar.

5.2.4 Semantic extraction

The topic modeling LDA, the sampling of Gibbs (Sethi et al., 2020; Blei et al., 2003), and the COVID-19-related analysis utilize lexical extraction and hidden topic discoveries. Furthermore, COVID-19 observations may vary on various topics. Recognize such significant topics and reveal them in this step. The collected documents, such as COVID-19, relevant notes, and words, were considered topics (K) based on the LDA model. Asymmetric Dirichlet distribution is generated from the distinct topic distributions. A COVID-19-related statement in a dataset has been used to measure the likelihood of observed values D.

ML methods assist the medical diagnosis of suspected COVID-19 for studying language indicators from the verbal expressions of older adults. Emphasize here that the optimal trained model for the illness group prediction includes strong lexical and high n-gram morphological characteristics (Figs. 2–4).

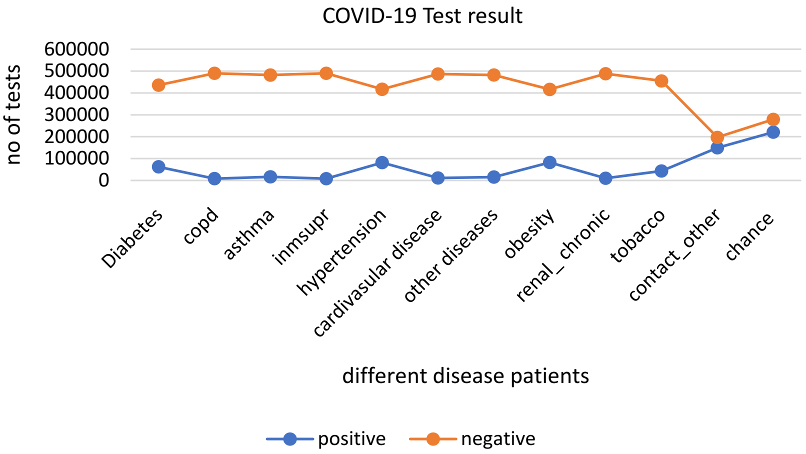

Fig. 2 Chance of COVID-19 test result of a patient with pre-existing diseases.

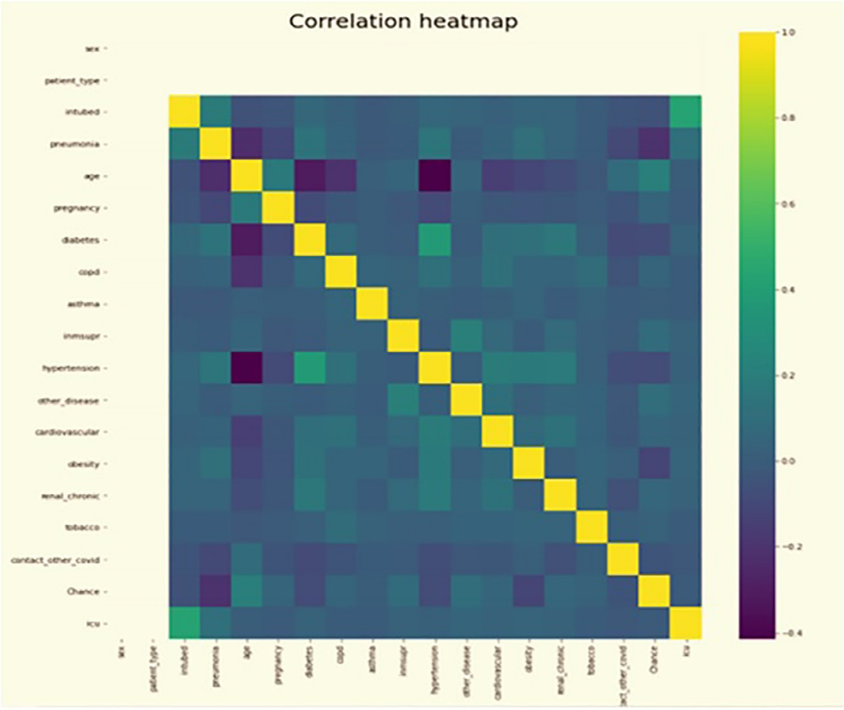

Fig. 3 Correlation heatmap.

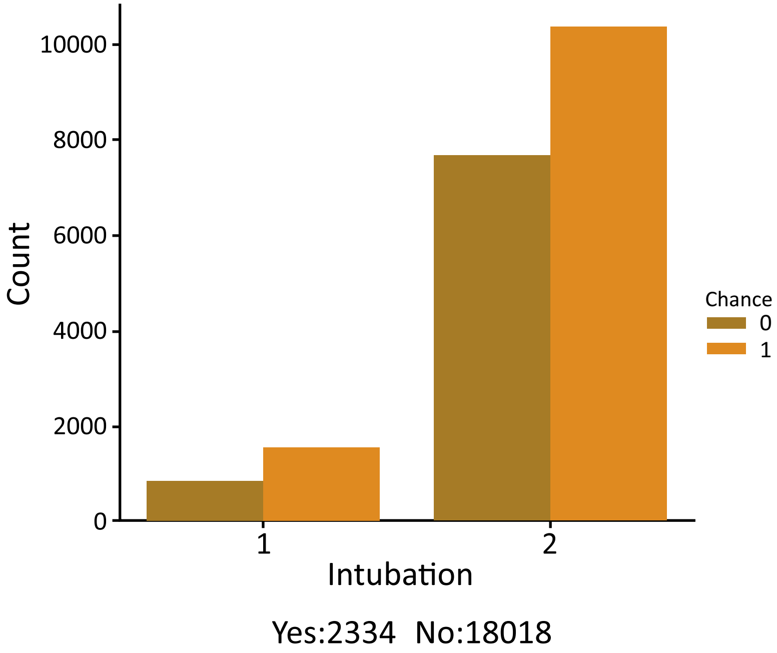

Fig. 4 Intubation COVID-19 patients. SVM classification accuracy: 64%.

Table 1 The COVID-19 patients with preexisting diseases.

	Sex	Patient_ type	Intubed	Pneumonia	Age	Pregnancy	Diabetes	Copd	Asthma	Inmsupr
0	2	1	97	2	27	97	2	2	2	2
1	2	1	97	2	24	97	2	2	2	2
2	1	2	2	2	54	2	2	2	2	2
3	2	2	2	1	30	97	2	2	2	2
4	1	2	2	2	60	2	1	2	2	2

Table 2 COVID-19 result with other chronic diseases.

S. No.	COVID-19 result	Positive	Negative
1	Diabetes	62,349	435,702
2	Copd	8276	489,970
3	Asthma	16,214	482,036
4	Inmsupr	8071	489,959
5	Hypertension	81,340	416,863
6	Cardiovascular disease	11,419	486,764
7	Other diseases	15,392	482,107
8	Obesity	81,929	416,293
9	Renal_ chronic	10,019	488,197
10	Tobacco	42,955	455,158
11	Contact_other	149,051	196,966
12	Chance	220,657	279,035

The following observations were completed by Tables 1 and 2. Topics 85 and 18 shared comparable notions of patients. Topic 85 included words referring to people, such as "people," "virus,", "day," "bad," "stop," "news," "worse," "sick," "spread," and "family." This topic is the first-ranked topic discovered from the generated latent topics, in which most users express their opinion and comment on this issue. Based on Table 1 and Fig. 5 in this topic, the terms people and virus were the most highlighted words, with word-weights of 0.1295% and 0.0301%, respectively. Also, we can see the importance of the term "family" from this topic. In addition, Topic 18 contains the telling words "virus," "people," "symptoms," "infection," "cases," "disease," "pneumonia," "coronavirus," and" treatment." Other revealing words in Topic 18 included "people," "infection," and "treatment." These terms initially suggest a set of user comments about treatment issues. Moreover, the sentiment analysis of the terms suggests that negative words were more highlighted than positive words (Tables 3 and 4).

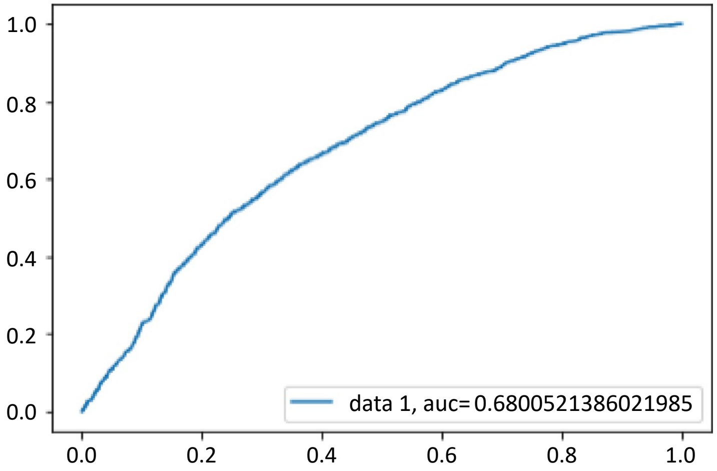

Fig. 5 Performance of COVID-19 tests receiver under area curve.

LM is trained to output a probability distribution over the vocabulary V given a context. Generically, p(V | c) = LM(c). Each word is typically encoding as a vector—a word embedding—learned as training and static across contexts. These word embeddings are stored in an input matrix Wi (n × |V|, where n is the embedding size (Figs. 6 and 7).

COVID-19 results

2—279,035, 1—220,657, 3—66,910

Where 2 represents COVID-19 negative, 1 represents COVID-19 positive, 3 represents COVID-19 waiting.

Now here we are replacing 1 with 1, 2 with 0, and 3 with 2.

This work is focusing on giving the chances of being affected by Coronavirus. So, the main focus is either chance is positive or negative, hence neglect awaiting cases.

6. Discussion

In this chapter, A COVID-19 pandemic model passed from humans to humans was designed and analyzed. To date, the suspected cases of COVID-19 continue to increase worldwide every day. Considering health risks, then, the forecast of infected people is especially notable. SVM classifies infected people as positive, negative, and waiting and predicting risk analysis those are suffering from underlying chronic diseases, such as diabetes, malignancy, and overweight. LDA models require the input parameters used, and it is challenging to choose this number since the correct number depends on the continuous measurement.

Table 3 LDA Gibbs 8 PD topic probabilities.

S. No.	V1	V2	V3	V4	V5	V6	V7	V8
1	0.115625	0.128125	0.140625	0.140625	0.103125	0.103125	0.140625	0.128125
2	0.135246	0.135246	0.118852	0.151639	0.118852	0.118852	0.102459	0.118852
3	0.128472	0.114583	0.142361	0.114583	0.086806	0.170139	0.128472	0.114583
4	0.090659	0.101648	0.123626	0.156593	0.090659	0.145604	0.156593	0.134615
5	0.109043	0.140957	0.119681	0.119681	0.119681	0.109043	0.119681	0.162234
6	0.204128	0.075688	0.094037	0.084862	0.12156	0.158257	0.167431	0.094037
7	0.144231	0.092949	0.105769	0.182692	0.080128	0.13141	0.13141	0.13141
8	0.11	0.136667	0.096667	0.136667	0.176667	0.15	0.11	0.083333
9	0.132143	0.103571	0.117857	0.146429	0.146429	0.117857	0.117857	0.117857
10	0.117188	0.127604	0.169271	0.075521	0.117188	0.117188	0.179688	0.096354
11	0.085616	0.167808	0.113014	0.126712	0.126712	0.099315	0.15411	0.126712
12	0.198052	0.107143	0.107143	0.12013	0.172078	0.107143	0.094156	0.094156
13	0.160714	0.089286	0.160714	0.089286	0.117857	0.103571	0.132143	0.146429
14	0.16358	0.175926	0.138889	0.101852	0.07716	0.114198	0.114198	0.114198
15	0.129747	0.129747	0.129747	0.117089	0.091772	0.167722	0.117089	0.117089
16	0.090625	0.115625	0.278125	0.090625	0.090625	0.090625	0.103125	0.140625
17	0.085294	0.144118	0.132353	0.108824	0.120588	0.167647	0.120588	0.120588
18	0.103125	0.128125	0.103125	0.178125	0.090625	0.090625	0.090625	0.215625
19	0.117089	0.129747	0.10443	0.155063	0.129747	0.167722	0.091772	0.10443
20	0.107558	0.130814	0.177326	0.177326	0.107558	0.130814	0.072674	0.09593

Table 4 LDA Gibbs 8 PD topics to terms.

S. No.	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
1	Breathing	Increase	COVID	Start	Recover	Exercises	Patients	Rest
2	Stress	Needs	Care	Health	Taken	Body	Physical	Best
3	Loved	Recovery	Health	Treatment	Pulmonary	People	Stay	Sore
4	Throat	Activity	Strong	Mental	Fear	Rehabilitation	Foods	Heart
5	Scared	Follow	Conditions	Keep	Importance	Drinking	Getting	Exercising
6	Back	Overcome	Advice	Fluids	Panic	Muscle	Negative	Isolation

Fig. 6 Word cloud based on Coronavirus.

Fig. 7 Word cloud based on Coronavirus.

The main issue of employing LDA models is an unsupervised training method; inadequate monitoring is often acceptable and expanded the test to examine if the semantic characteristics of diagnostic tests may depend on the different COVID-19-related concerns. In this context, an existing dataset considered significant Latin subjects of COVID-19 keywords was linked to numerous difficulties. LDA is a probabilistic model that implies that every word in a dataset creates a mixture of (latent) unnoticed subjects applied to documents. A topic is described as a categorical word distribution. Therefore, it is concluded that there are prospects of quality management if more information is collected with these machine learning algorithms. As we have a significant issue in addressing the deadly virus, the work benefits the population in a certain sense by analyzing diagnostic tests and necessary action. Also, it was examined that people with preexisting diseases with COVID-19 have predicted risk analysis.

7. Conclusion

The COVID-19 pandemic highlighted the significance of managing a medical crisis in diagnostic procedures that significantly affect the global social, financial, and healthcare. In this work, the diagnosis of COVID-19 using SVM has three limitations: selecting a feature vector, an effective SVM input feature subset, and establishing the best kernel parameters. These limitations are significant, since choosing the subset feature influences the kernel parameters and likewise. The number of samples in developing assessment methods is an important aspect. Along with the strategic evaluation of an uncertainties associated lexicon, the word vector was designing. In addition, a sample of feelings senses is considering. The added annotation makes it possible to differentiate emotional sensations with intensity and multifactorial. Lexical Semantically Analysis depicted the vector illustration emotions classifications and realized a practical weight function wherein a generic phrase's compelling and valuable meaning was defined. The vocabulary of feeling in investigations is of enormous importance. A good quality lexicon guarantees the efficiency of the process supervised.

8. Future work

Fuzzy methods will be used in the future to effectively analyze the content of pharmaceutical assessments using phrase vectors and dynamic classification algorithms. These subsequent tests are supporting for intensive reasons. It is necessary to identify the generative models that create theoretical and realistic knowledge for COVID-19 diagnostic analyses. Predicting preclinical diseases, pandemic trends, and precise analyses of human-to-human spread in medical personnel and serosurveys. Though subsequent diagnoses are necessary for managing and controlling outbreaks, fast, scalable, and high-precision research has the most significant cost-efficient multiplexing diagnostics that identify several diseases at once.

References

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., Shah, Z., 2020. Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. J. Med. Internet Res. 22 (4), e19016.

Acebal, J., Banchs, R., Crego, J., Brosa, A., Lambert, P., Fonollosa, J., Costa-Jussa, M., 2006. N-gram-based machine translation. Comput. Linguist. 32 (4), 527–549. https://doi.org/10.1162/coli.2006.32.4.527.

Ahmed, H., Traore, I., Saad, S., 2017. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. pp. 127–138.

Al-A'abed, M., Al-Ayyoub, M., 2016. A lexicon-based approach for emotion analysis of arabic social media content. In: Proceedings of the International Computer Sciences and Informatics Conference (ICSIC 2016), Amman, Jordan.

Bakshi, R.K., Kaur, N., Kaur, R., Kaur, G., 2016. Opinion mining and sentiment analysis. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, pp. 452–455.

Bernheim, A., Mei, X., Huang, M., Yang, Y., Fayad, Z.A., Zhang, N., Diao, K., Lin, B., Zhu, X., Li, K., Li, S., Shan, H., Jacobi, A., Chung, M., 2020. Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection. Radiology 70. https://doi.org/10.1148/radiol.2020200463, 200463.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Boleda, G., 2020. Distributional semantics and linguistic theory. Annu. Rev. Linguist. 6, 213–234.

Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., Hassanien, A.E., 2020. Sentiment analysis of COVID-19 tweets by deep learning classifiers-a study to show how popularity is affecting accuracy in social media. Appl. Soft Comput. 97, 106754.

Chan, J.F.-W., Yip, C.C.-Y., Tang, T.H.-C., Wong, S.C.-Y., Leung, K.-H., Fung, A. Y.-F., Ng, A.C.-K., Zou, Z., Tsoi, H.-W., Choi, G.K.-Y., Tam, A.R., Cheng, V. C.-C., Chan, K.-H., Tsang, O.T.-Y., Yuen, K.-Y., 2020. Improved molecular diagnosis of COVID-19 by the novel, highly sensitive, and specific COVID-19-RdRp/Hel real-time reverse transcription-PCR assay validated in vitro and with clinical specimens. J. Clin. Microbiol. 58 (5).

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297.

Dahal, B., Kumar, S.A.P., Li, Z., 2019. Topic modeling and sentiment analysis of global climate change tweets. Social Network. Anal. Min., 1–20. https://doi.org/10.1007/s13278-019-0568-8.

Deep Learning Demystified, 2021. https://deeplearningdemystified.com/article/nlp-1. (Accessed 18 May 2021).

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. J. Am. Soc. Inf. Sci. 41 (6), 391–407.

Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., Makhoul, J., 2014. Fast and robust neural network joint models for statistical machine translation. In: Proceedings of the

Association for Computational Linguistics (ACL), Baltimore, pp. 1370–1380, https://doi.org/10.3115/v1/P14-1129.

Drus, Z., Khalid, H., 2019. Sentiment analysis in social media and its application: systematic literature review. Prog. Comput. Sci. 161, 707–714.

European Commission, 2020. Current Performance of COVID-19 Test Methods and Devices and Proposed Performance Criteria. EC, Brussels. https://www.mayoclinic.org/tests-procedures/covid-19-diagnostic-test/about/pac-20488900. (Accessed 12 May 2021).

Falkum, I., Vicente, A., 2015. Polysemy: current perspectives and approaches. Lingua 157, 1–172. https://doi.org/10.1016/j.lingua.2015.02.002.

Frege, G., 1967. Begrifsschrif, a formula language, modelled upon that of arithmetic, for pure thought. In: From Frege to Godel: A Source Book in Mathematical Logic, 1879–1931, pp. 1–82.

García-Díaz, J.A., Cánovas-García, M., Valencia-García, R., 2020. Ontology-driven aspect-based sentiment analysis classification: an infodemiological case study regarding infectious diseases in Latin America. Futur. Gener. Comput. Syst. 112, 641–657.

Guo, W.L., Jiang, Q., Ye, F., Li, S.Q., Hong, C., Chen, L.Y., Li, S.Y., 2020. Effect of throat washings on detection of 2019 novel coronavirus. Clin. Infect. Dis. 71 (8), 1980–1981. https://doi.org/10.1093/CID/ciaa416. 32271374.

Hai Son, L., Allauzen, A., Yvon, F., 2012. Continuous space translation models with neural networks. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies, pp. 39–48.

Jelodar, H., Wang, Y., Orji, R., Huang, H., 2020. Deep sentiment classification and topic discovery on novel Coronavirus or covid-19 online discussions: NLP using LSTM recurrent neural network approach. IEEE J. Biomed. Health Inform. 24, 2733–2742.

Kaila, R.P., Prasad, A.V.K., 2020. Informational flow on twitter – coronavirus outbreak – topic modeling approach. Int. J. Adv. Res. Eng. Technol. 11 (3), 128–134.

Kakulapati, V., Reddy, S.M., 2020. Risk analysis of coronavirus caused death by the probability of patients suffering from chronic diseases – a machine learning perspective. JCR 7 (14), 2626–2633.

Kakulapati, V., Reddy, S.M., Deepthi, B.S.S., Tavares, J.M.R.S., 2020. Machine learning analysis of topic modeling re-ranking of clinical records. In: Advances in Ubiquitous Sensing Applications for Healthcare, Smart Biosensors in Medical Care. Academic Press, ISBN: 9780128207819, pp. 153–177. ISSN 25891014.

Kevadiya, B.D., Machhi, J., Herskovitz, J., et al., 2021. Diagnostics for SARS-CoV-2 infections. Nat. Mater. 20, 593–605.

Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. 104 (2), 211–240.

Levelt, W.J.M., 2008. Formal Grammars in Linguistics and Psycholinguistics: Volume 1: An Introduction to the Theory of Formal Languages and Automata, Volume 2: Applications in Linguistic Theory; Volume 3: Psycholinguistic Applications. John Benjamins Publishing.

Liu, S., Pan, C., 2020. Differentiating diagnosis of COVID-19 or influenza in patients based on laboratory data during flu season. EClinicalMedicine 26, 100511.

Mardian, Y., Kosasih, H., Karyana, M., Neal, A., Lau, C.Y., 2021. Review of current COVID-19 diagnostics and opportunities for further development. Front. Med. 8, 615099.

Medhat, W., Hassan, A., Korashy, H., 2014. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. 5 (4), 1093–1113.

Mikolov, T., Yih, W.-t., Zweig, G., 2013. Linguistic regularities in continuous space word representations. In: Proceedings of HLT-NAACL, pp. 746–751.

Mukhaiyar, U., Pasaribu, U., 2012. A new procedure for generalized STAR modeling using the IAcM approach. ITB J. Sci. 44, 179–192. https://doi.org/10.5614/itbj.sci.2012.44.2.7.

Muthusami, R., Bharathi, A., Saritha, K., 2020. COVID-19 outbreak: tweet-based analysis and visualization towards the influence of coronavirus in the world. Gedrag Organ. Rev. 33, 534–549.

Nurhayati, N., Pasaribu, U., Neswan, O., 2012. Application of generalized space-time autoregressive model on GDP data in West European countries. J. Prob. Stat., 1–16. https://doi.org/10.1155/2012/867056.

Pan, Y., Zhang, D., Yang, P., Poon, L.L.M., 2020. Viral load of SARS-CoV-2 in clinical samples. Lancet Infect. Dis. 20 (4), 411–412.

Ravi, K., Ravi, V., 2015. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl.-Based Syst. 89, 14–46.

Read, J.M., Bridgen, J.R.E., Cummings, D.A.T., Ho, A., Jewell, C.P., 2020. Novel coronavirus 2019-nCoV: early estimation of epidemiological parameters and epidemic predictions. medRxiv 2020. https://doi.org/10.1101/2020.01.23.20018549, 20018549.

Rivers, C., Chretien, J.P., Riley, S., et al., 2019. Using "outbreak science" to strengthen the use of models during epidemics. Nat. Commun. 10, 3102.

Sanyaolu, A., Okorie, C., Marinkovic, A., Ayodele, O., Abbasi, A.F., Prakash, S., Ahmed, M., Kayode, D., Jaferi, U., Haider, N., 2020. Navigating the diagnostics of COVID-19. SN Compr. Clin. Med., 1–8.

Schwenk, H., 2012. Continuous space translation models for phrase-based statistical machine translation. In: COLING (Posters), pp. 1071–1080.

Sethi, M., Pandey, S., Trar, P., Soni, P., 2020. Sentiment Identification in COVID-19 Specific Tweets. pp. 509–516.

Van Loon, N., Verbrugghe, M., Cartuyvels, R., Ramaekers, D., 2021. Diagnosis of COVID-19 based on symptomatic analysis of hospital healthcare Workers in Belgium: observational study in a large Belgian tertiary care center during early COVID-19 outbreak. J. Occup. Environ. Med. 63 (1), 27–31.

Verma, P., Khanday, A., Rabani, S., Mir, M., Jamwal, S., 2019. Twitter sentiment analysis on Indian government project using R. Int. J Recent Technol. Eng. 8, 8338–8341. https://doi.org/10.35940/ijrte.C6612.098319.

Whiting, P., Singatullina, N., Rosser, J.H., 2015. Computed tomography of the chest: I. basic principles. Contin. Educ. Anaesth. Crit. Care Pain 15 (6), 299–304. https://doi.org/10.1093/bjaceaccp/mku063.

World Health Organization, 2019. https://www.who.int/. (Accessed 26 April 2021).

Wu, F., Zhao, S., Yu, B., et al., 2020a. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269.

Wu, J.T., Leung, K., Leung, G.M., 2020b. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. Lancet 395 (10225), 689–697. Epub 2020 Jan 31. Erratum in: lancet.

Xu, W., Liu, X., Gong, Y., 2003. Document Clustering Based On Non-negative Matrix Factorization. SIGIR Forum (ACM Special Interest Group on Information Retrieval), pp. 267–273, https://doi.org/10.1145/860435.860485.

Yang, Z., Zeng, Z., Wang, K., Wong, S.S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z., Mai, Z., Liang, J., Liu, X., Li, S., Li, Y., Ye, F., Guan, W., Yang, Y., Li, F., Luo, S., Xie, Y., Liu, B., Wang, Z., Zhang, S., Wang, Y., Zhong, N., He, J., 2020. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J. Thorac. Dis. 12 (3), 165–174.

Zhang, L., Wang, S., Liu, B., 2018. Deep learning for sentiment analysis: a survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 8 (4), e1253.

Further reading

Corman, V.M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D.K., Bleicker, T., Brünink, S., Schneider, J., Schmidt, M.L., Mulders, D.G., Haagmans, B.L., van der Veer, B., van den Brink, S., Wijsman, L., Goderski, G., Romette, J.L., Ellis, J., Zambon, M., Peiris, M., et al., 2020. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro. Surveill. 25 (3), 2000045.

Jelodar, H., Wang, Y., Yuan, C., et al., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed. Tools Appl. 78, 15169–15211.

Loeffelholz, M.J., Tang, Y.W., 2020. Laboratory diagnosis of emerging human coronavirus infections – state of the art. Emerg. Microbes Infect. 9 (1), 747–756.

Mimno, D., Wallach, H.M., McCallum, A., 2008. Gibbs sampling for logistic normal topic models with graph-based priors. In: NIPS Workshop on Analyzing Graphs. vol. 61.

Prabhakar Kaila, D., Prasad, D.A., et al., 2020. Informational flow on twitter-corona virus outbreak-topic modelling approach. Int. J. Adv. Res. Eng. Technol. 11 (3), 128–134.

Streiner, D.L., Cairney, J., 2007. What's under the ROC? An introduction to receiver operating characteristics curves. Can. J. Psychiatr. 52 (2), 121–128.