Manchester Metropolitan University

e-space

Manchester Metropolitan University's Research Repository

Please cite the Published Version

Sampath, Pradeepa, Packiriswamy, Gayathiri, Pradeep Kumar, Nishmitha, Shanmuganathan, Vimal, Song, Oh-Young, Tariq, Usman and Nawaz, Raheel (2020) IoT Based Health—Related Topic Recognition from Emerging Online Health Community (Med Help) Using Machine Learning Technique. Electronics, 9 (9). p. 1469.

DOI: https://doi.org/10.3390/electronics9091469

Publisher: MDPI AG

Version: Published Version

Downloaded from: https://e-space.mmu.ac.uk/626505/

Usage rights: Creative Commons: Attribution 4.0

Additional Information: MDPI, copyright The Author(s).

Enquiries:
If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines)

electronics

MDPI

Article

IoT Based Health—Related Topic Recognition from Emerging Online Health Community (Med Help) Using Machine Learning Technique

Pradeepa Sampath [1], Gayathiri Packiriswamy [1], Nishmitha Pradeep Kumar [1], Vimal Shanmuganathan [2], Oh-Young Song [3,*], Usman Tariq [4] and Raheel Nawaz [5]

1 School of Computing, SASTRA Deemed to Be University, Tirumalaisamudram, Thanjavur 613401, Tamil Nadu, India; pradeepa@it.sastra.edu (P.S.); gayuswamy2998@gmail.com (G.P.); nishupradeep98@gmail.com (N.P.K.)
2 Department of IT, National Engineering College, Kovilpatti, Thoothukudi District 628503, India; svimalphd@gmail.com
3 Department of Software, Sejong University, Seoul 05006, Korea
4 College of Computer Engineering and Science, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia; u.tariq@psau.edu.sa
5 Department of Operations, Technology, Events and Hospitality Management, Manchester Metropolitan University, Manchester M1 5GD, UK; R.Nawaz@mmu.ac.uk
* Correspondence: oysong@sejong.edu

Received: 9 July 2020; Accepted: 1 September 2020; Published: 8 September 2020

Abstract: The unprompted patient's and inimitable physician's experience shared on online health communities (OHCs) contain a wealth of unexploited knowledge. Med Help and eHealth are some of the online health communities offering new insights and solutions to all health issues. Diabetes mellitus (DM), thyroid disorders and tuberculosis (TB) are chronic diseases increasing rapidly every year. As part of the project described in this article comments related to the diseases from Med Help were collected. The comments contain the patient and doctor discussions in an unstructured format. The sematic vision of the internet of things (IoT) plays a vital role in organizing the collected data. We pre-processed the data using standard natural language processing techniques and extracted the essential features of the words using the chi-squared test. After preprocessing the documents, we clustered them using the K-means++ algorithm, which is a popular centroid-based unsupervised iterative machine learning algorithm. A generative probabilistic model (LDA) was used to identify the essential topic in each cluster. This type of framework will empower the patients and doctors to identify the similarity and dissimilarity about the various diseases and important keywords among the diseases in the form of symptoms, medical tests and habits.

Keywords: online health community; diabetes; tuberculosis; thyroid; chi-squared test; K-means++; LDA; IoT; machine learning technique

1. Introduction

The metabolic disease diabetes mellitus, the contagious infection tuberculosis and thyroid disease are major chronic diseases which affect billions of people every year. These chronic diseases rapidly increased death rates over past decades and they act as a gateway to several other diseases by weakening the immune system of humans. According to the World Health Organization, 422 million people are affected by diabetes and 1.6 million deaths occur each year due to diabetes and tuberculosis [1]. A BioMed Centre (BMC) public health journal survey indicates that lower levels of thyroid hormones increase the risk of diabetes mellitus.

Diabetes mellitus is a metabolic disease in which blood glucose levels are divergently high. Insulin is a hormone produced by the pancreas and is responsible for lowering the glucose level in blood. Insufficient production of insulin, absence of insulin and an inability of human bodies to properly utilize insulin are major causes of diabetes [2]. Diabetes mellitus is categorized as type1 or insulin-dependent or juvenile-onset diabetes and type2 or insulin-independent or adult-onset diabetes [3]. In the United States, diabetes is the seventh most common cause for death.

Tuberculosis is an infectious disease caused by a bacterium called Mycobacterium tuberculosis (MTB). Tuberculosis (TB) directly affects lungs and also invades through other organs. It spreads from one person to another person through coughs, sneezes and saliva. TB is categorized into active TB or extrapulmonary TB and latent TB infection. The BCG vaccine acts as a barrier to the deadly disease tuberculosis. The WHO describes TB as an "epidemic" and proclaims that tuberculosis is one of the preeminent causes of death by a single contagious agent [4].

The thyroid gland is a butterfly-shaped endocrine gland present in the neck. The thyroid gland is responsible for producing thyroid hormones that control various metabolic activities in the human body. An abnormal increase or decrease of the thyroid hormone leads to thyroid disease. Thyroid disease is classified into hyperthyroidism, or an overactive thyroid, and hypothyroidism, which is an underactive thyroid. Hashimoto disease, Graves' disease, thyroid nodules and goiter are the most prominent disorders of the thyroid. Thyroid disease is a truculent disease, which is almost impossible to eradicate and exists in the human body throughout its lifetime [5].

Social media platforms support reciprocated computing-mediated technologies that facilitate users to share new information, ideas and their opinions [6] with their communities. Online health communities (OHCs) and health care professionals (HCPs) are an emerging phenomenon in social media which connect various groups of individuals having similar health-related issues and interests [7–11]. Using this persuasive platform HCPs clarify public health-related problems, illustrate the use of health care policy and practice issues, promote public health programs, motivate patients and educate every individual by providing continuous support and service.

The information collected from Med Help, e-Health, WebMD, Healthline, Medscape, Everyday Health and Health Central are helpful in identifying inter-relationships among generally arising acute diseases [12]. The keywords from the collected information are helpful for patients and physicians to explore information about these chronic diseases. The knowledge gathered from these keywords acts as an aegis to reduce the possible death rate.

The analysis of 750 messages collected for four different chronic diseases depicts a perception of a diverse and varied range of activities carried out by moderators [13,14]. Community development and a strengthening of local networks help to improve the quality of life (QOL) of older people and self-harming behavior patients affected by various diseases [15–17].

In the work [18] data are collected from a Zambia rural community and the analyzed results evidently explain the experience and responsibility of the mother, who satisfies cultural and health expectation during new-born care. Through community content and thematic relationships, the effect of climatic changes on human physical and mental health are explained in [19]. Text mining and science mapping techniques are used to analyze and interpret the results [20]. A systematic pharmacological method is combined with other data mining techniques for the evaluation of drug similarity [21].

Dataset data mining techniques are applied on a dataset of MTA (Metropolitan Transportation Authority) customer feedback to enhance QOS (quality of service) and identify customer satisfaction levels [22]. The tool interprets and identifies diagnostics patterns from a huge free clinical dataset of notes of patients, using text mining techniques [23]. The study used the K-means++ algorithm to increase accuracy of the recommendation system [24].

An improved K-means algorithm and dimensionality reduction were used to perform clustering of Arabic text [25]. A K-means text clustering algorithm was efficiently used in spam detection [26]. In another study a weighted K-means algorithm text clustering was performed [27].

IoT and cloud services are playing a major role for extracting and visualizing the data without human intervention [28–32].

The analysis reports of trusted health care organizations are an important source from which to find relationships between the mentioned chronic diseases. The online health community platform is recommended by physicians to obtain accurate knowledge about all diseases. The OHC texts play a vital role in the extraction of keywords and in finding inter-relationships between all three chronic diseases.

The objectives are designed in a way to emphasize social values and to eradicate lingering diseases, namely diabetes mellitus (DM), tuberculosis (TB) and thyroid disorders. The three prominent objectives are delineated as follows:

- To extract important keywords of each disease from each cluster.
- To find inter–relationships among three chronic diseases.
- To measure the accuracy of extracted keywords by comparing keywords with the world's trusted organization reports.

2. Materials and Methods

Figure 1 portrays the overall architecture of the system. The key points to highlight regarding our contribution to the proposal are listed below:

1. The comments discussed by both patients and physicians in the healthcare forum, Med Help, are collected for the three chronic diseases.
2. The datasets are pre-processed using NLP techniques such as tokenization, stopword removal and punctuation removal. The term frequency-inverse document frequency (TF–IDF) measure is used to collect the most important words from the collected pre-processed datasets.
3. The most important feature words are filtered using the chi-square test from the three pre-processed datasets.
4. The K-means++ algorithm is applied to the reduced feature datasets. With evidence of clustering groups, LDA is used to identify the most frequently occurring meaningful keywords.
5. Keywords identified from each cluster of all three diseases are compared with the world's trusted healthcare organizations to measure their accuracy.

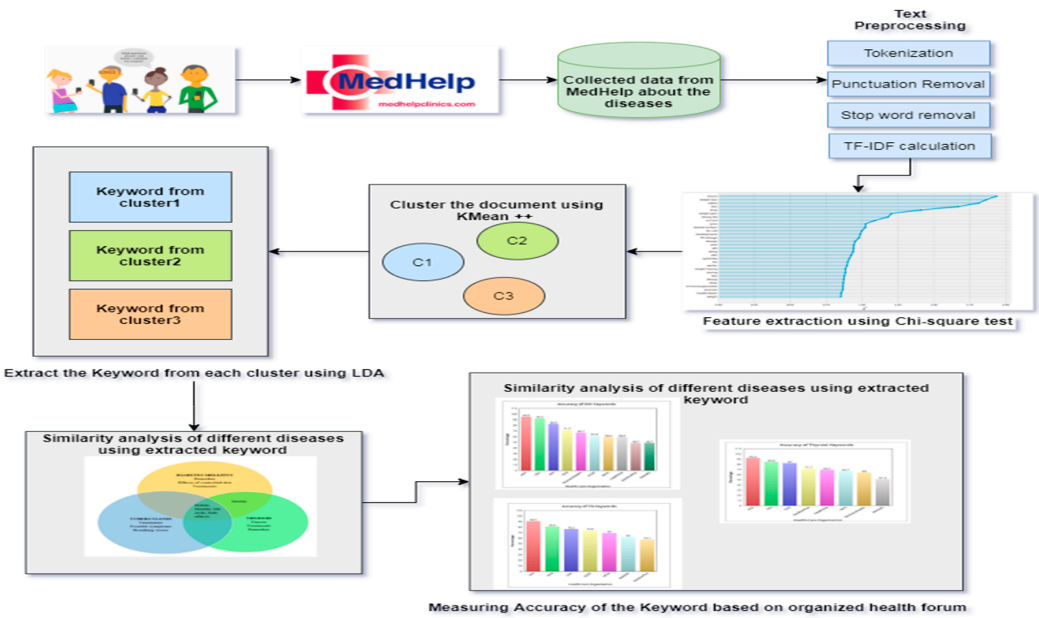


Figure 1. Overall Architecture.

2.1. Data Set Gathering and Preprocessing

In this study, the online health community, Med Help, is used as a platform to collect the dataset. Powerful web API's are helpful in the translation of information in the connected world (IoT) [10]. The precautions, remedies and knowledge about the three chronic diseases are discussed in comments by physicians and patients in the online health community [33]. The comments discussed are stored in the Med Help cloud for numerous diseases and disorders. The comments discussed about diseases are collected as a dataset over the years of 2018, 2019 and 2020 (up to January) using a web API; the results are then stored in a local database.

In NLP, pre-processing is an inevitable step where normal texts are transformed into a simple form. Pre-processing is an underlying step responsible for better performance of machine learning (ML) algorithms. Tokenization is a pre-processing step where paragraphs are split into sentences and sentences are split into individual words. Stop words are connecting words in a sentence which do not produce intent meaning. Stopwords are removed in the pre-processing step by utilizing a manually created stopword dictionary or prebuilt libraries based on sensitivity. Term frequency-inverse document frequency, shortly known as TF–IDF [34], is an important statistical measure which calculates the importance of a word in a document or in a corpus.

The term frequency-inverse document frequency (TF–IDF) calculation is described here. The TF value of a term "t" in a document "d" is given by the frequency "f" of that term in the document, divided by the number of words in that document, as mentioned in Equation (1). The IDF value for a word refers to its importance within the whole dataset, considering its occurrence in every document, as given in Equation (2). The TF–IDF value is merely the product of these values, represented in Equation (3). Algorithm for text preprocessing is discussed in Algorithm 1.

$$\text{TF}\ (t, d) = \left(\frac{\text{f}_{(t,d)}}{\text{number of words in } d} \right) \tag{1}$$

$$\text{IDF}\ (t) = \ln \left(\frac{n_{\text{documents}}}{n_{\text{documents containing } t}} \right) \tag{2}$$

$$\text{TF-IDF}\left(t_k, d_j\right) = \text{TF(word)} \times \text{IDF(word)} \tag{3}$$

Algorithm 1 Text-Preprocessing

Input: Dataset Collected from Medhelp.
Output: Pre-processed dataset for each disease

1. Until all words in the document in a file are exhausted
   - Tokenization
   - Stemming
   - Punctuation removal
   - Stopword Removal
2. Calculate the TF value from Equation (1)
3. Calculate the inverse document frequency from Equation (2)
4. Calculate the TF–IDF value, set a minimum threshold value using Equation (3)
5. If TF-IDF score >the threshold value (0.53)
   - Append the word into a document
6. End

2.2. Chi-Square Test

Algorithm 2 portrays the execution of the chi-square Test. Feature selection or attribute selection is a process of extracting the most relevant features from a dataset. The chi-square test is a statistical measure used for feature selection based on the dependency of words in a document. The chi-square test is calculated using the following formula:

$$X^2 = \frac{(\text{Observed Frequency of words} - \text{Expected Frequency of words})2}{\text{Expected Frequency of words}} \tag{4}$$

where

- Observed frequency is the number of observations of words in a document,
- Expected frequency is the number of expected observations of words in a document if there is no relationship between features.

Algorithm 2 Chi-Square Test

Input: Pre-processed dataset
Output: Essential feature dataset, F extracted based on Chi-Square test.

```
1.  Dataset D = d1 . . .  dn.
2.  d1 = w1 . . .  wn.
3.  For d in D
4.  For w in d

    calculate the expected frequency e
    calculate the observed frequency o
    score = (squares of observed and expected)/observed
    if(score<threshold)//threshold = 0.47
    F = F + w
5.  End for
6.  End for
7.  End
```

2.3. K-Mean++

K-means++ is an unsupervised iterative clustering algorithm. Algorithm 3 discusses the pseudo code for K-Mean++ techniques. K-means ++ is an extended version of the popular K-means algorithm, which ensures perfect and nimble initialization of centroids and enhances the quality of clustering [35]. The K-means++ algorithm is limited to numerical values and groups similar documents into a single group in a corpus. Doc2Vec is a deep learning unsupervised algorithm which generates feature vectors for documents. The generated feature vectors are used to find similarity between documents.

Algorithm 3 K-means++

Input: Essential feature dataset extracted based on the chi-square test.
Output: Seven clustered documents.

1. Data points d = d1, d2 . . . dn.
2. Choose one center k randomly from the data points d.
3. Initial centroid C1 = k.
4. For x in d
   - Find the nearest centroid (C2 . . . Cn) using the distance formula
   - Assign cluster Cj = x
5. Selection of next centroid is based on the probability that relies upon the distance of the first initialized centroid
6. Repeat the steps (4) and (5) until all centroids (C1 . . . Ck) have been sampled.
7. End

2.4. LDA

LDA is an unsupervised machine learning model used for topic modelling [30]. In LDA, each document is considered as a topic mixture and each topic is considered as a mixture of words. Several words describe the same topic and several topics construct the same document. LDA represents the correct meaning of words in topic modelling as compared to LSA (latent semantic analysis). LDA provides better results and accuracy than LSA [36,37]. Spectral clustering is used to cluster the document and PNN classifier is used to identify the label of the cluster discussed in [38]. Different classifier algorithms and machine learning techniques are used to classify the data set and text document [39–42]. LDA and LSA are used to identify the topics from the given text document without clustering.

3. Results

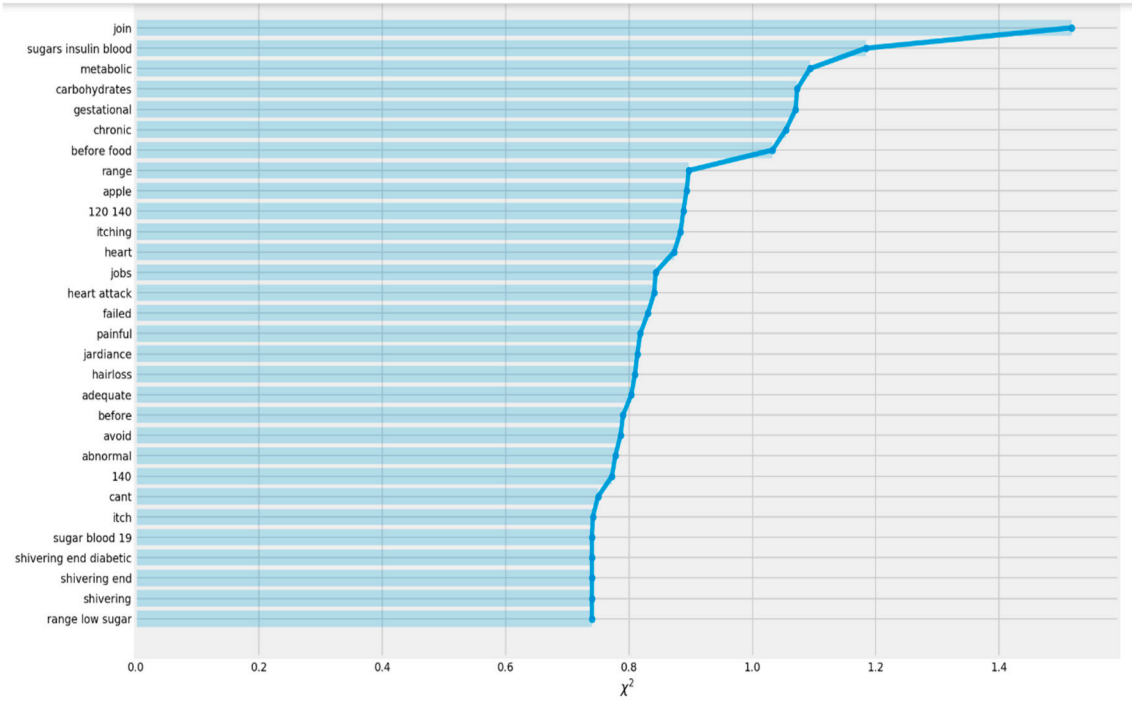
3.1. Data Set Gatering and Preprocessing

The dataset of diabetes mellitus consists of 74, 233 and 311 documents from 2020, 2019 and 2018, respectively. The tuberculosis dataset consists of 625 documents, which comprise 117, 276 and 232 documents from 2020, 2019 and 2018, respectively. The thyroid dataset is composed of 591 documents, which contain 116, 219 and 256 documents from 2020, 2019 and 2018, respectively. From all three diseases, a total of1824 documents are collected from the online health community.

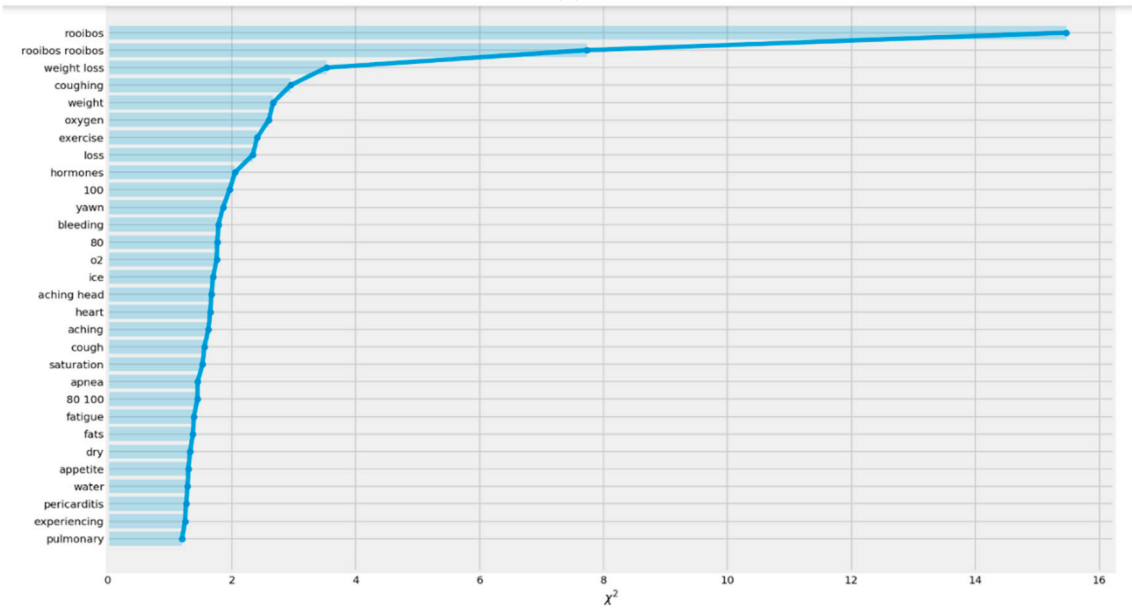
The collected dataset of each disease is pre-processed using the Python NLP NLTK and Scikit-learn packages, which include tokenization, stop word removal and punctuation removal methods. The TF–IDF measure is used to collect the most important words in a dataset and to remove low-frequency terms from the corpus. After the pre-processing step, the pre-processed dataset of each disease contains the most meaningful words.

3.2. Chi-Square Test

The chi-square test is applied to each pre-processed dataset to extract the most important features based on a threshold value. The extracted important features are recorded. Figure 2 shows the results of the Chi-square test for each dataset. In this figure the top 30 words of each disease are represented in graph format based on the threshold value of 0.47.

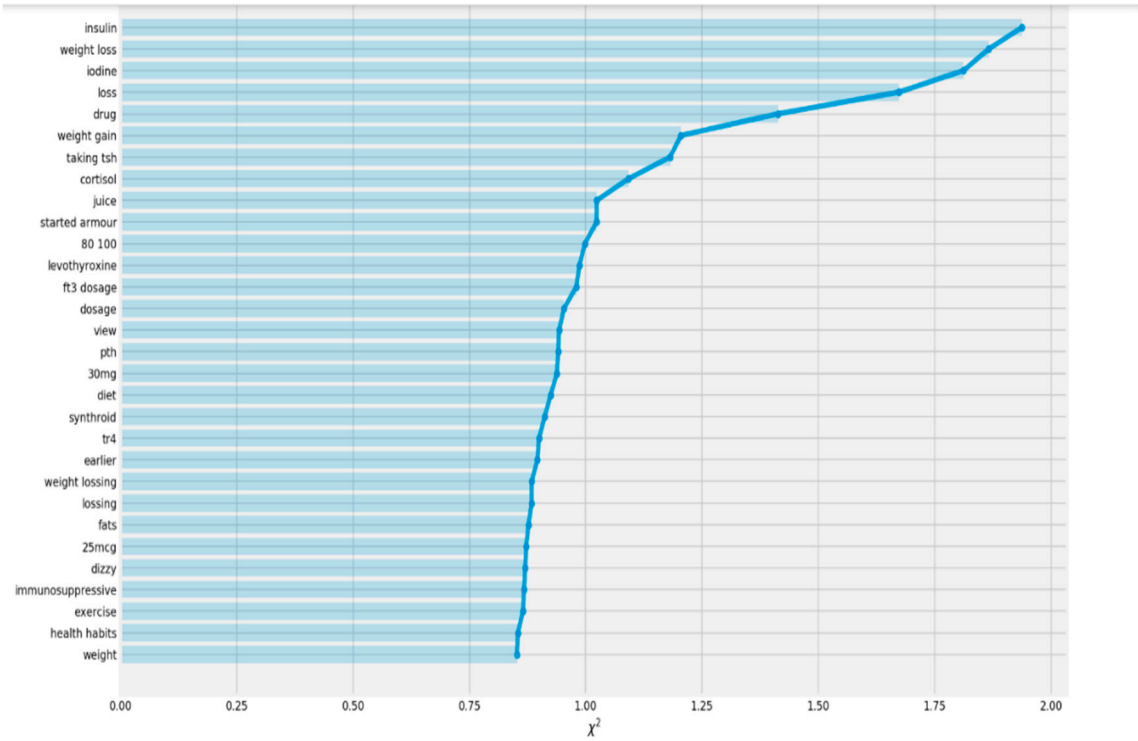


(**a**)



(**b**)

Figure 2. *Cont.*

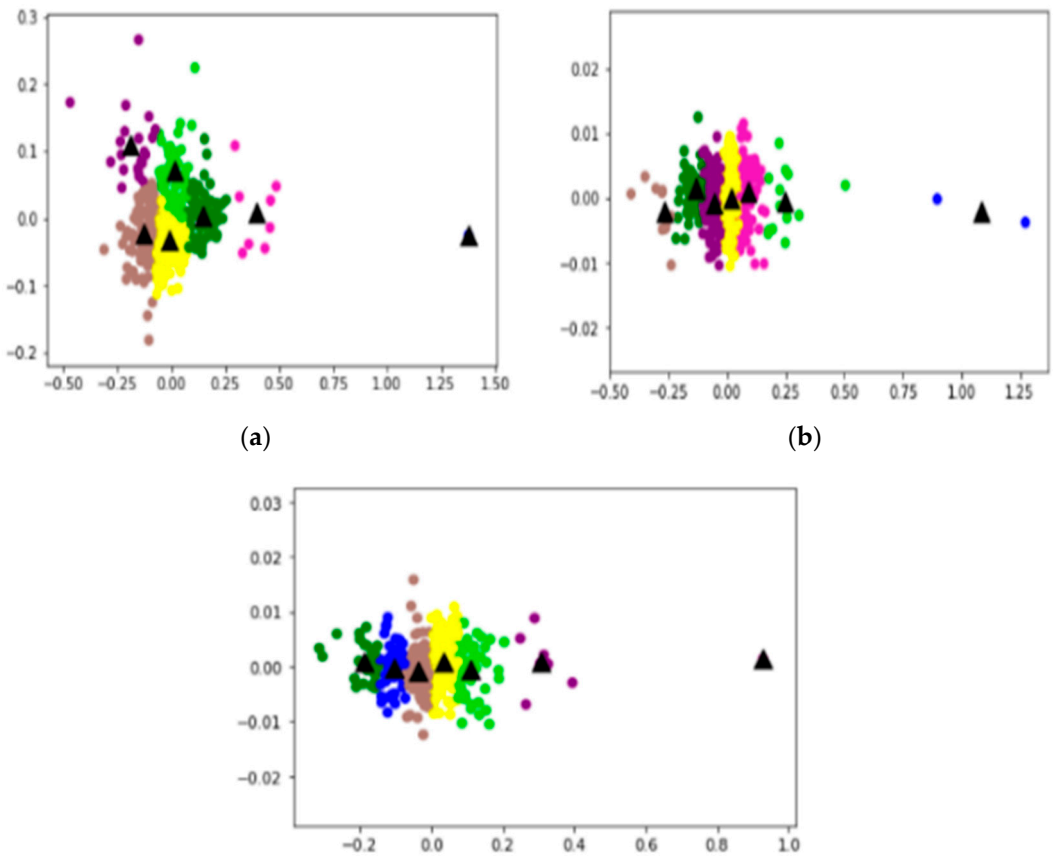


(**c**)

Figure 2. The chi-square test depicts the top 30 words: (**a**) DM, (**b**) thyroid and (**c**) TB.

3.3. K-Mean++

The Python Gensim package includes Doc2Vec.Whenutilizing the Gensim package, Doc2Vec is applied to the datasets which are collected based on the chi-square test for each disease. Doc2Vec generates feature vector values based on the similarity between documents. The K-means++ algorithm is applied to the dataset which is retrieved based on the Doc2Vec feature vector values for each disease. For each disease, the clustering process is repeated with seven clusters. Each cluster document is collected individually for all diseases. The clusters of the three diseases are depicted in Figure 3.



(**a**) (**b**)

(**c**)

Figure 3. K-means++ clustering of: (**a**) DM, (**b**) TB and (**c**) thyroid.

3.4. LDA

LDA is applied to each cluster of all three diseases to retrieve the top ten topics. As a result of LDA, the most important keywords are extracted and each cluster is manually labelled based on the keywords for all three chronic diseases. Word cloud is a visualization technique used to visualize high-frequency terms in each cluster for all diseases. Count vectorizer is used to visualize the most frequently occurring words of each cluster of all three diseases in Figure 4.

(**a**) (**b**) (**c**)

Figure 4. Visualization of keywords byword cloud: (**a**) DM, (**b**) TB and (**c**) thyroid.

4. Discussion

The clusters are labelled manually for all three diseases. The most prominent keywords of each cluster are tabulated for all three diseases. The most important keywords are extracted as a result of the LDA process for all three diseases. The sample terms of each cluster are extracted based on the sample terms inter-relationships between all three diseases. Authors should discuss the results and how they can be interpreted in perspective of previous studies and of the working hypotheses. The findings and

their implications should be discussed in the broadest context possible. Future research directions may also be highlighted. Keywords about the diseases are listed in the Tables 1–3.

Table 1. Sample keywords extracted from each cluster of thyroid.

Clusters of Thyroids	Sample Terms
Remedies	Treats ferritin level, maintain iodine level, hormone gland test, eltroxin pills
Side effects	Weight problems, fatigue, blood pressure, sleep apnea, iron deficiency, hair loss, heart problems, dizzy
Habits	Exercise, diet, healthy food
Treatments	Ft3, tsh, ft4 blood test, iodine free, vitamin b12, hormone balancing, treat hashimoto, ferritin level
Insulin	Insulin level, hypothyroid, blood sugar level, vitamin and protein level, hormone problem
Healthy Lifestyle	Hormone balance, exercise, sleep, steroids
causes	Hormone imbalance, auto-immune system disorder, hashimoto, constipation, goiter, high calcium consumption, stress

Table 2. Sample keywords extracted from each cluster of diabetes mellitus (DM).

Clusters of DM	Sample Terms
Remedies	Hair loss Victoza, diet, control hormone, glucose test, 120–140, low carbs, avoid high carbs, healthy foods
Side effects	Weight problems, fatigue, blood pressure, sleep apnea, iron deficiency, hair loss, heart problems, dizzy
Habits	Exercise, diet, healthy food
Treatments	Hpa1c test, controlling blood pressure, low carbs fasting, treat auto immune system, aerobics, exercise, metformin pills
Insulin	Insulin level, hypothyroid, blood sugar level, vitamin and protein level, hormone problem
Healthy Lifestyle	Hormone balance, exercise, sleep, steroids
Effects of Controlled diet	Reduce heart risk, lower urine infection, maintaining insulin level, treats hypoglycemia

Table 3. Sample keywords extracted from each cluster of tuberculosis (TB).

Clusters of Tuberculosis	Sample Terms
Remedies	Nexium pills, treating migraine, visit psychiatrist, ultrasound scan, yoga, inhalers
Side effects	Weight problems, fatigue, blood pressure, sleep apnea, iron deficiency, hair loss, heart problems, dizzy
Habits	Exercise, diet, healthy food
Treatments	Inhalers, avoid liquids, treat sleep apnea, pulmonary test, drink water, treats chronic
Healthy Lifestyle	Hormone balance, exercise, sleep, steroids
Possible symptoms	Headache, cough, breathing problem, stomach problem, fungal infections, appetite, chest pain
Breathing issues	Smoking, drug, cold water, cold drinks, high blood pressure, lack of sleep, weight, allergy

Side effects, Habits and Healthy Lifestyle are the clusters which are found common in all three diseases. Based on this inference, the relationship between the three chronic diseases is found.

Side effects: In the Side effects cluster problems faced by each disease patient are grouped. The interpretation is that the patients of all three diseases are facing common health problems even though the cause of all three diseases is different.

Habits: The Habits cluster demonstrates that pre-activity should be carried out by patients to prevent all three chronic diseases.

Healthy Lifestyle: The Healthy Lifestyle cluster describes the activities that should be carried out by patients to recover from diseases and to prevent death.

Side effects, Habits, and Healthy Lifestyle are three clusters which were found common among all three chronic diseases. These three clusters and their respective keywords evidently depict the prominent inter-relationships between diabetes mellitus, tuberculosis and thyroid disease. Venn diagrams are used to analyze common themes among all three diseases. A Venn diagram interprets common themes among diabetes mellitus, tuberculosis and thyroid disease and also illustrates similarity among diabetes mellitus and thyroid disease. Side effects, Habits and Healthy Lifestyle are common themes between all three chronic diseases, which are found from Venn diagram interpretation. It is represented in Figure 5. The common themes identified among the three chronic diseases reveal an occurrence of inter-relationship between them. The cause and impact of the three chronic diseases are different but the cluster similarity among the three diseases evidently describes inter-relationships between the three diseases.

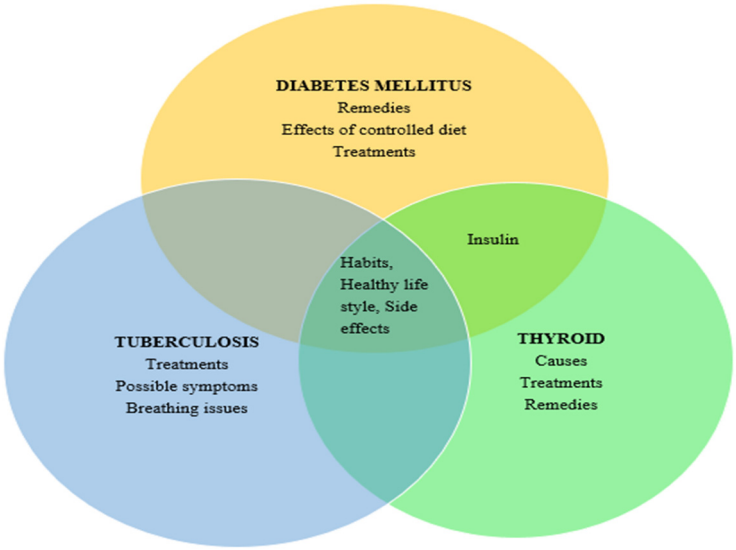


Figure 5. A Venn diagram summarization to depict inter-relationships between the three chronic diseases.

The accuracy score of a keyword is measured based on the number of keywords extracted and is mapped with the world's trusted organization reports. Keywords of each cluster extracted from all diseases are compared with the world's trusted organization reports. The comparison results illustrate accuracy of each keyword of all clusters, which evidently shows the accuracy of each keyword. The World Health Organization (WHO), the National Health Survey (NHS), the National Institute of Health (NIH), the Centre for Disease Control and Prevention (CDC), the European Centre for Disease Control and Prevention (ECDC), the National Centre for Disease Control and Prevention (NCDC), the American Diabetes Association (ADA), the American Thyroid Association (ATA), Women's Health, MedlinePlus, WebM and Healthline are twelve of the world's trusted organizations. The mentioned twelve organization reports are compared to measure accuracy of all keywords for all diseases. Accuracy scores of each cluster keyword, compared with trusted organization reports, are tabulated.

The comparison result evidently illustrates that each keyword of all clusters extracted from all disease datasets are accurate and they can be interpreted to have a factual meaning. The sample

keywords from each cluster are compared with the mentioned 12 organization reports. Based on the occurrence of the keywords, each cluster accuracy is measured in percentage and the results are tabulated in Tables 4–6.

Table 4. Comparison of DM keywords with different healthcare organization.

Organization/Clusters of DM Accuracy in %	Remedies	Side Effects	Habits	Treatments	Insulin	Healthy Lifestyle	Effects of Controlled Diet	Overall Accuracy in %
WHO	62.5	50	100	42.8	60	50	75	58.9
CDC	87.5	100	100	85.7	100	75	100	92.3
Women's Health	75	62.5	100	57.1	60	50	75	66.7
NHS	75	62.5	100	42.8	80	100	75	71.7
NCDC	62.5	50	100	57.1	60	50	75	61.5
ADA	87.5	100	100	85.7	100	100	100	94.8
NIH	75	75	100	85.7	80	75	100	82.5
MedlinePlus	50	37.5	66.6	57.1	40	25	75	48.7
Healthline	62.5	50	100	71.4	40	50	50	58.9
WebMD	50	50	66.6	42.8	60	50	75	48.7

Table 5. Comparison of thyroid keywords with different healthcare organization.

Organization/Clusters of Thyroid Accuracy in %	Treatments	Side Effects	Insulin	Remedies	Healthy Lifestyle	Habits	Causes	Overall Accuracy in %
CDC	87.5	75	80	100	75	100	71.4	82
Women's Health	62.5	75	60	75	50	66.7	57.1	64
NHS	75	75	60	75	50	66.7	57.1	66.7
ATA	100	100	80	100	75	100	85.7	92.3
NIH	87.5	87.5	80	75	75	100	85.7	84.6
MedlinePlus	100	25	60	100	75	66.6	85.7	71.7
Healthline	87.5	75	60	25	75	66.6	71.4	69.2
WebMD	50	37.5	60	50	75	66.6	42.8	51.3

Table 6. Comparison of TB keywords with different healthcare organization.

Organization/Clusters of TB Accuracy in %	Side Effects	Possible Symptoms	Habits	Treatments	Remedy	Breathing Issues	Healthy Lifestyle	Overall Accuracy in %
CDC	62.5	57.1	66.7	83.3	83.3	75	50	69
Women's Health	75	85.7	33.3	83.3	66.7	87.5	75	76.2
NHS	62.5	100	66.7	100	83.3	75	75	80.9
ATA	62.5	57.1	66.7	66.7	83.3	87.5	100	73.8
NIH	100	85.7	66.7	83.3	100	87.5	100	90.5
MedlinePlus	62.5	71.4	66.7	66.7	50	75	25	62
Healthline	62.5	71.4	66.7	66.7	66.7	25	25	57.1
WebMD	75	71.4	33.3	33.3	83.3	75	75	66.7

The American Diabetes Association gives 94.8% accuracy for keywords of the disease diabetes mellitus (DM). The majority of diabetes mellitus keywords are matched with ADA reports. The keywords of thyroid are well mapped with the American Thyroid Association reports, which in turn produce 92.3% overall accuracy. The tuberculosis sample keywords are majorly matched with the National Institute of Health reports, which show an overall accuracy of 90.5%.

5. Conclusions

This framework is helpful for general users and patients to obtain knowledge about all three chronic diseases in the form of causes, medications, side effects and remedies. Lack of effective analysis tools to discover hidden relationships and trends among these diseases led us to propose a model that made use of technological advancements in text mining to develop a prediction, detection and treatment model for the chronic diseases problem. The consummated analysis and findings reduce the burden of physicians and encourage various physicians to conduct numerous health programs, which create tremendous awareness among the society. The dataset is collected for only three years from

2018 to 2020. The collection of a larger number of documents would help to extract more keywords for all three diseases. The datasets are collected from only one online health community platform-Med Help. The datasets collected from different online health community platforms would help to analyze more documents. The analysis will be helpful to extract more keywords accurately. Analysis of other dependent diseases helps to reduce the rate of death caused by the perilous diseases. Further analysis will create awareness among people and will reduce the death rate enormously.

Author Contributions: Data curation, P.S., G.P., N.P.K., V.S., O.-Y.S., U.T. and R.N.; formal analysis, P.S., G.P., N.P.K. and V.S.; funding acquisition, O.-Y.S.; investigation, O.-Y.S., U.T. and R.N.; methodology, V.S. and R.N.; supervision, V.S., U.T. and R.N.; validation, O.-Y.S., U.T. and R.N.; visualization, O.-Y.S., U.T. and R.N.; writing—original draft, P.S., G.P., N.P.K. and V.S.; writing—review and editing, P.S., G.P. and N.P.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2016-0-00312) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and in part by the MSIP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW)(2015-0-00938) supervised by the IITP(Institute for Information & communications Technology Planning&Evaluation).

Conflicts of Interest: The authors declare that they do not have any conflict of interests. This research does not involve any human or animal participation. All authors have checked and agreed with the submission.

References

1. Vrieling, F.; Ronacher, K.; Kleynhans, L.; van den Akker, E.; Walzl, G.; Ottenhoff, T.H.; Joosten, S.A. Patients with Concurrent Tuberculosis and Diabetes have a Pro-Atherogenic Plasma Lipid Profile. *EbioMedicine* **2018**, *32*, 192–200. [CrossRef] [PubMed]
2. Fiarni, C.; Sipayung, E.M.; Maemunah, S. Analysis and Prediction of Diabetes Complication Disease using Data Mining Algorithm. *Procedia Comput. Sci.* **2019**, *161*, 449–457. [CrossRef]
3. Wang, Y.; Yang, H.; Huynh, Q.; Nolan, M.; Negishi, K.; Marwick, T.H. Diagnosis of Nonischaemic Stage B Heart Failure in Type 2 Diabetes Mellitus: Optimal Parameters for Prediction of Heart Failure. *JACC Cardiovasc. Imaging* **2018**, *11*, 1390–1400. [CrossRef] [PubMed]
4. Su, F.C.; Friesen, M.C.; Humann, M.; Stefaniak, A.B.; Stanton, M.L.; Liang, X.; LeBouf, R.F.; Henneberger, P.K.; Virji, M.A. Clustering asthma symptoms and cleaning and disinfecting activities and evaluating their associations among healthcare workers. *Int. J. Hyg. Environ. Health* **2019**, *222*, 873–883. [CrossRef] [PubMed]
5. dos Santos, B.S.; Steiner, M.T.; Fenerich, A.T.; Lima, R.H. Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. *Comput. Ind. Eng.* **2019**, *138*, 106120. [CrossRef]
6. Nilashi, M.; Ibrahim, O.; Yadegaridehkordi, E.; Samad, S.; Akbari, E.; Alizadeh, A. Travelers decision making using online review in social network sites: A case on Trip Advisor. *J. Comput. Sci.* **2019**, *28*, 168–179. [CrossRef]
7. Jia, S.S. Motivation and satisfaction of Chinese and U.S. tourists in restaurants: A cross-cultural text mining of online reviews. *Tour. Manag.* **2020**, *78*, 104071. [CrossRef]
8. Lenzi, A.; Maranghi, M.; Stilo, G.; Velardi, P. The social phenotype: Extracting a patient-centered perspective of diabetes from health-related blogs. *Artif. Intell. Med.* **2019**, *101*, 101727. [CrossRef]
9. Zhou, J.; Zuo, M.; Ye, C. Understanding the factors influencing health professionals' online voluntary behaviors: Evidence from YiXinli, a Chinese online health community for mental health. *Int. J. Med. Inform.* **2019**, *130*, 103939. [CrossRef]
10. Introne, J.; Goggins, S. Advice reification, learning and emergent collective intelligence in online health support communities. *Comput. Hum. Behav.* **2019**, *99*, 205–218. [CrossRef]
11. Zhang, Y.; Ibaraki, M.; Schwartz, F.W. Disease surveillance using online news: Dengue and zika in tropical countries. *J. Biomed. Inform.* **2020**, *102*, 103374. [CrossRef] [PubMed]
12. Park, A.; Conway, M.; Chen, A.T. Examining thematic similarity, difference and membership in three online mental health communities from reddit: A text mining and visualization approach. *Comput. Hum. Behav.* **2018**, *78*, 98–112. [CrossRef] [PubMed]
13. Smedley, R.M.; Coulson, N.S. A thematic analysis of messages posted by moderators within health-related asynchronous online support forums. *Patient Educ. Couns.* **2017**, *9*, 1688–1693. [CrossRef] [PubMed]

14. Hewison, A.; Atkin, K.; McCaughan, D.; Roman, E.; Smith, A.; Smith, G.; Howell, D. Experiences of living with chronic myeloid leukemia and adhering to tyrosine kinase inhibitors: A thematic synthesis of qualitative studies. *Int. J. Nurs. Sci.* **2020**, *6*, 50–57.
15. Nuntaboot, K.; Boonsawasdgulchai, P.; Bubpa, N. Roles of mutual help of local community networks in community health activities: Improvement for the quality of life of older people in Thailand. *Int. J. Nurs. Sci.* **2019**, *6*, 266–271. [CrossRef]
16. Stoltenberg, D.; Maier, D.; Waldherr, A. Community detection in civil society online networks: Theoretical guide and empirical assessment. *Soc. Netw.* **2019**, *59*, 120–133. [CrossRef]
17. Leung, M.; Chow, C.B.; Ip, P.K.; Yip, S.F. Self-harm attempters' perception of community services and its implication on service provision. *Int. J. Nurs. Sci.* **2019**, *6*, 50–57. [CrossRef]
18. Lovell, N.; Etkind, S.N.; Bajwah, S.; Maddocks, M.; Higginson, I.J. Control and Context Are Central for People with Advanced Illness Experiencing Breathlessness: A Systematic Review and Thematic Synthesis. *J. Pain Symptom Manag.* **2019**, *57*, 140–155. [CrossRef]
19. Buser, J.M.; Moyer, C.A.; Boyd, C.J.; Zulu, D.; Ngoma-Hazemba, A.; Mtenje, J.T.; Jones, A.D.; Lori, J.R. Cultural beliefs and health-seeking practices: Rural Zambians' views on maternal-newborn care. *Midwifery* **2020**, *85*, 102686. [CrossRef]
20. Moro, A.; Joanny, G.; Moretti, C. Emerging technologies in the renewable energy sector: A comparison of expert review with a text mining software. *Futures* **2020**, *117*, 102511. [CrossRef]
21. Zhang, C.; Zhao, Y.; Zhang, X. An improved association rule mining-based method for discovering abnormal operation patterns of HVAC systems. *Energy Procedia* **2019**, *158*, 2701–2706. [CrossRef]
22. Ghazzawi, A.; Alharbi, B. Analysis of Customer Complaints Data using Data Mining Techniques. *Procedia Comput. Sci.* **2019**, *163*, 62–69. [CrossRef]
23. Ribeiro, J.; Duarte, J.; Portela, F.; Santos, M.F. Automatically detect diagnostic patterns based on clinical notes through Text Mining. *Procedia Comput. Sci.* **2019**, *160*, 684–689. [CrossRef]
24. Song, Y.T.; Wu, S. Slope One Recommendation Algorithm Based on User Clustering and Scoring Preferences. *Procedia Comput. Sci.* **2020**, *166*, 539–545. [CrossRef]
25. Sangaiah, A.K.; Fakhry, A.E.; Abdel-Basset, M.; El-henawy, I. Arabic text clustering using improved clustering algorithms with dimensionality reduction. *Clust. Comput.* **2019**, *22*, 4535–4549. [CrossRef]
26. Sasaki, M.; Shinnou, H. Spam Detection Using Text Clustering. In Proceedings of the International Conference on Cyberworlds, Singapore, 23–25 November 2005.
27. Chen, X.; Yin, W.; Tu, P.; Zhang, H. Weighted k-means Algorithm Based Text Clustering. In Proceedings of the International Symposium on Information Engineering and Electronic Commerce, Ternopil, Ukraine, 16–17 May 2009.
28. Wang, C.; Yang, G.; Papanastasiou, G.; Zhang, H.; Rodrigues, J.; Albuquerque, V. Industrial Cyber-Physical Systems-based Cloud IoT Edge for Federated Heterogeneous Distillation. *IEEE Trans. Ind. Inform.* **2020**. [CrossRef]
29. Wang, C.; Dong, S.; Zhao, X.; Papanastasiou, G.; Zhang, H.; Yang, G. Saliencygan: Deep learning semisupervised salient object detection in the fog of IoT. *IEEE Trans. Ind. Inform.* **2019**, *16*, 2667–2676. [CrossRef]
30. Annamalai, S.; Udendhran, R.; Vimal, S. An Intelligent Grid Network Based on Cloud Computing Infrastructures. *Nov. Pract. Trends Grid Cloud Comput.* **2019**, 59–73. [CrossRef]
31. Annamalai, S.; Udendhran, R.; Vimal, S. Cloud-Based Predictive Maintenance and Machine Monitoring for Intelligent Manufacturing for Automobile Industry. *Nov. Pract. Trends Grid Cloud Comput.* **2019**, 74–81. [CrossRef]
32. Shafiq, M.; Tian, Z.; Bashir, A.K.; Du, X.; Guizani, M. CorrAUC: A Malicious Bot-IoT Traffic Detection Method in IoT Network Using Machine Learning Techniques. *IEEE Internet Things J.* **2020**, *132*, 1. [CrossRef]
33. Can, U.; Alatas, B. A new direction in social network analysis: Online social network analysis problems and applications. *Phys. A Stat. Mech. Appl.* **2019**, *5351*, 122372. [CrossRef]
34. Rashid, J.; Shah, S.M.; Irtaza, A.; Mahmood, T.; Nisar, M.W.; Shafiq, M.; Gardezi, A. Topic Modelling technique for text Mining Over Biomedical Text Corpora Through Hybrid Inverse Documents Frequency and Fuzzy K-means Clustering. *IEEE Access* **2019**, *7*, 146070–146080. [CrossRef]
35. Vargas-Calderón, V.; Camargo, J.E. Characterization of citizens using word2vec and latent topic analysis in a large set of tweets. *Cities* **2019**, *92*, 187–196. [CrossRef]

36. Yang, S.; Huang, G.; Cai, B. Discovering Topic Representative terms for Short Text Clustering. *IEEE Access* **2019**, *9*, 92037–92047. [CrossRef]
37. Momtazi, S. Unsupervised Latent Dirichlet Allocation for supervised question classification. *Inf. Process. Manag.* **2018**, *54*, 380–393. [CrossRef]
38. Pradeepa, S.; Manjula, K.R.; Vimal, S.; Khan, M.S.; Chilamkurti, N.; Luhach, A.K. DRFS: Detecting Risk Factor of Stroke Disease from Social Media Using Machine Learning Techniques. *Neural Process. Lett.* **2020**. [CrossRef]
39. Shafiq, M.; Tian, Z.; Bashir, A.K.; Jolfaei, A.; Yu, X. Data mining and machine learning methods for sustainable smart cities traffic classification: A survey. *Sustain. Cities Soc.* **2020**, *60*, 23. [CrossRef]
40. Geetha, R.; Sivasubramanian, S.; Kaliappan, M.; Vimal, S.; Annamalai, S. Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier. *J. Med. Syst.* **2019**, *43*, 286. [CrossRef]
41. Ramamurthy, M.; Krishnamurthi, I.; Vimal, S.; Robinson, Y.H. Deep learning based genome analysis and NGS-RNA LL identification with a novel hybrid model. *Biosystems* **2020**, *197*, 104211. [CrossRef]
42. Iweni, C.; Khan, S.; Anajemba, J.H.; Bashir, A.K.; Noor, F. Realizing an efficient IoMT-assisted Patient Diet Recommendation System through Machine Learning Model. *IEEE Access* **2020**, *8*, 28462–28474. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).