A Machine Learning Approach for Exploring CORD-19 Scholarly Articles

Subedi, Abhishek
Universität passau
Email: subedi01@ads.uni-passau.de

***Abstract*—The WHO has declared COVID-19 a pandemic that is caused by SARS-CoV-2 virus and has created challenges among scientist and engineers. Extensive research are going on to find possible vaccines and solutions to control the pandemic. Collaborations between scientists, research institutions are very vital for the success. They require fast and efficient access of information from huge amount of research articles published about covid, virus, pandemic and related topics.**

In response to the pandemic, the contribution of this research will be to apply ML and NLP techniques to explore insights from the resource of approximately 10000 scholarly articles about CORD-19.

***Index Terms*—Cord-19, scholarly articles, clustering, classification, search, cosine similarity**

I. Introduction

The WHO has declared COVID-19 a pandemic that is caused by SARS-CoV-2 virus. With increase of the diseases, researchers around the world are involved in extensive research to understand and find possible vaccines. Researchers in past and in present have published thousands of papers and articles related to SARS and its variants. [1].

They need to know about the research done by several other scientists. Hence from the vast amount of information, it is difficult for them to find the relevant and most important papers about particular topics in the domain. Discovery of knowledge and insight from data to understand the ongoing scenario is essential for scientist and policy makers in tackling the situations. In response to the pandemic, this research apply Machine Learning and Natural Language Techniques to explore insights from the resources of over ten thousand scholar articles about CORD-19.

II. Research Objective

As discussed, it is difficult to get close to important documents and information from huge amount of scholar articles. This research will explore the scholar datasets that will shorten the research time providing relevant information required for the scientists. The main focus of the paper are given below.

A. Clustering and Classification

The datasets used are unlabeled, so unsupervised k-mean algorithms will be used to classify the relevant articles into several clusters. The cluster value will be used as label to the data. To evaluate how well the cluster has generalized, a classifier will be trained to predict the cluster.

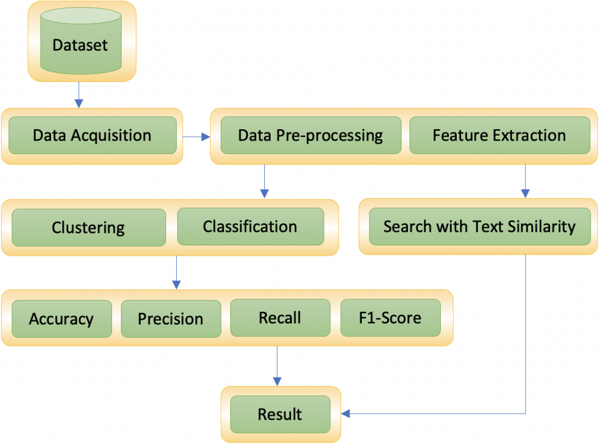


Fig. 1. Project Workflow

B. Visualisation

Interactive visualization of clusters will be done so as to give user freedom of exploring the information dynamically. Word cloud will be used in order to get insight about what each clusters are representing.

C. Search with Text Similarity

Scholar articles will be searched based on the text similarity, i.e., users inputs the query and Cosine similarity between scholarly articles and query will be computed to find the most relevant papers from whole datasets.

III. Workflow

The Figure 1, gives an overview of project workflow and the detail description of it are described as follow.

A. Data Acquisition

The acquired datasets of scholar articles about coronavirus are created by Allen Institute for AI and provided as a free resource for the global research community. The purpose of datasets is to apply ML and NLP to find new insights in support of fight against virus. The datasets have research articles related to COVID-19 and it's variants. [2]

B. Data Pre-processing

Since the extracted data are not well structured and clean to be used for further processing, pre-processing is required such that it is ready to use for input to the algorithms and

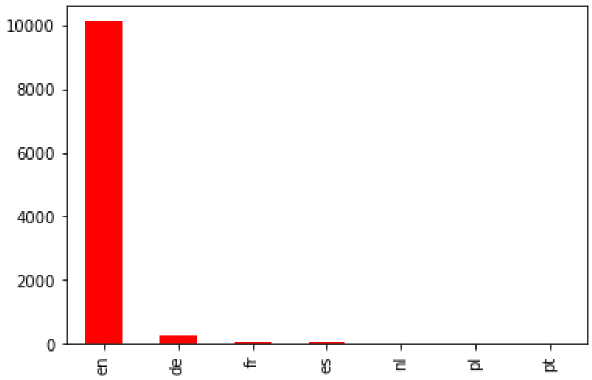


Fig. 2. Language Distribution

visualizations. As the size of datasets is very large approximately 25GB and contains more than 280,000 scholarly articles, this research will make use of less data (approx. 10000) considering the computational capacity. Therefore the following steps are done to pre-process data.

- The metadata of datasets has null values in the columns $title$, $abstract$ and pcm_json_files: file path to full article. Since this columns are important for feature extraction and ML models, the rows that does not have this information are discarded.
- The json files are parsed to get full literature and $title$, $abstract$, $full\_literature$ are combined to make a complete text of scholarly article.
- The whole datasets are converted to lower case.
- Punctuation's and non alphabetic characters are removed.
- Multi-space are reduced to single space
- Articles that are not in English language are removed. The Fig. 2, shows an overview of language distribution.
- Each articles in rows are tokenized so as to remove stop words, remove pronoun and finally combine into sentences .
- The articles that contains minimum 3000 and maximum 10000 words are selected for the research which accounts exactly 10126 scholarly articles. The Fig. 3, gives the distribution of word count though out the datasets. This distribution shows articles having word in between 3000 to 10000 are high in density.
- Finally the working datasets are exported to a .csv file format. Fig. 4 shows a subset of the datasets after performing pre-processing steps.

1) Tokenization: Tokenization is the process of splitting a given sentences or a document into pieces, called tokens.

C. Feature Extraction

Feature extraction is the task of transforming input data into a set of features, which are distinctive properties of input patterns that help in differentiating between the categories of input patterns [3]. The following feature extraction technique is used in the research.

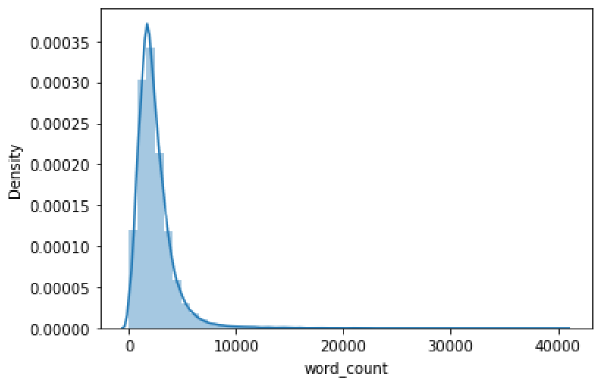
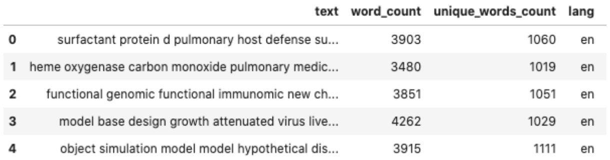


Fig. 3. Word Count Distribution



	text	word_count	unique_words_count	lang
0	surfactant protein d pulmonary host defense su...	3903	1060	en
1	heme oxygenase carbon monoxide pulmonary medic...	3480	1019	en
2	functional genomic functional immunomic new ch...	3851	1051	en
3	model base design growth attenuated virus live...	4262	1029	en
4	object simulation model model hypothetical dis...	3915	1111	en

Fig. 4. Few articles from the dataset

1) TF-IDF (Term Frequency-Inverse Document Frequency): This method is intended to reflect how important a word is to a document in a collection or corpus based on its frequency in the corpus. The Fig. 5, depicts importance of word based on occurrence.

In the datasets, some term will occur more frequently but will carry very little meaningful information. These very frequent an less meaningful terms will dim the frequencies of unique yet more interesting terms. TF-IDF intend to scale down the impact of such words.

Inverse Document Frequency is a numerical measure of how

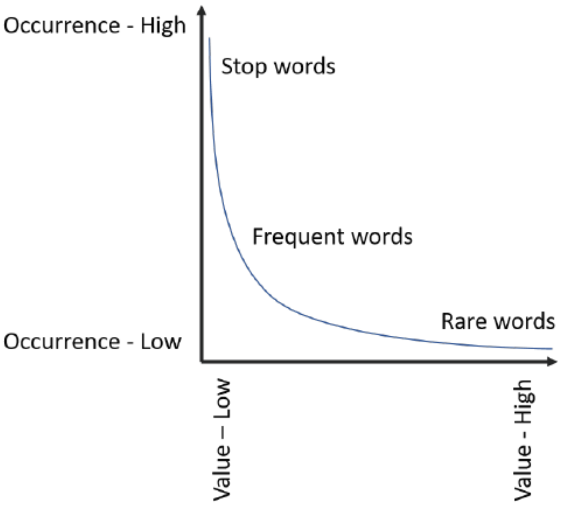


Fig. 5. Occurrence of word and it's importance

much information a term provides [4]:

$$IDF(t, D) = log \frac{|D|}{DF(t, D)}$$

Where t denotes a term, D denotes the corpus. $DF(t, D)$ is the document frequency, the number of documents that contains term t.

The TF-IDF is simply the product of TF and IDF [3]:

$$TFIDF(t, d, D) = TF(t, d) \cdot IDF(t, D)$$

Where, if a term appears in all documents, its IDF value becomes 0 and $TF(t, d)$ is Term Frequency, the number of times that term t appers in document d.

D. Dimensionality Reduction

Large data are hard to explain, hence a technique that reduces the dimension which helps to interpret data with minimum information loss is required. The research applies Principal Component Analysis (PCA) for dimensionality reduction [5].

1) Principal Component Analysis (PCA): PCA follows the following methods for reducing the dimension of the datasets.

- *Normalization :* The first step in PCA is to normalize the datsets by subtracting the respective means for each datasets.

$$\bar{x} \Leftarrow \frac{1}{n} \sum_{i=1}^{n} (x_i)$$

$$X \Leftarrow subtract\ \bar{x}\ from\ each\ row\ x_i\ in\ X$$

- *Covariance Matrix Computation:* Compute the covariance of normalized data. The covariance of two variable is a measure of their correlation. On the other hand correlation shows how strongly two variables are related to each other.

$$COV \Leftarrow \frac{1}{n-1} (X)^T * (X)$$

- *Compute Eigenvectors and correspoinding Eigenvalues:* Next calculate eigenvectors and values for covariance matrix. In general the eigenvector of a matrix A is the vector for which the following holds:

$$A\vec{v} = \lambda(\vec{v})$$

  where, λ is the eigenvalue that is scalar in nature.
- *Rank eigenvalue from largest to smallest:* After computing eigenvalues of covariance matrix, rank the eigenvectors w.r.t. decreasing order of eigenvalues. Hence, get first k eigenvectors called feature vector which will be the dimension of new datasets.
- *Forming Principal Component:* The last step is to build new reduced data from the k chosen matrix of vectors as given below.

$$New\ Data = FeatureVector \cdot (NormalizedData)^T$$

  Where FeatureVector is the first k eigenvectors acquired by ranking eigenvalues in descending order. Normalized data are obtained from the first step in PCA as mentioned above.

Hence going through the algorithm and applying to the research, PCA successfully decreased the dimension of scholarly articles from (10126, 5000) to (10126, 2292) with minimum information loss.

E. Machine Learning Task

This phase will be completed in following order:

- Use Elbow method to find best possible number of clusters
- Run KMeans with best cluster value
- Visualize the cluster and word cloud
- Run classification to evaluate how well the clusters has generalized.
- Search scholarly article with Text Similarity.

1) Clustering: As the datasets are unlabeled, a simple unsupervised machine learning algorithm called K-means is used. It groups the articles into K clusters, where K is assigned by user. To get a better cluster the right number of cluster must be assigned. Therefore, Elbow method is used to find the right number of clusters.

The general approach is to run k-means clustering for value of $2 < K < 50$. Hence for each cluster find the sum of squared errors.

- ***Cluster Distortion***: It gives the squared distance from the cluster center for particular cluster. In general Euclidean distance is applied to calculate the squared distance [6].

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

  Where p, q are two points in Euclidean space, q_i, p_i are Euclidean vectors that initiates from origin and n is the total space.

The fig. 6, visualize the result of elbow method. For this research $K = 3$ is used as number of clusters.

- ***K-Means***: Since the datasets are unlabeled the best approach is to group data which are similar to one another known as clusters. K-Means is widely used clustering algorithm. It uses K centroids to define clusters. A data point belongs to a cluster's centroid if it is near to that centroid than other centroids [5]. The **Algorithm 1** shows KMeans clustering algorithm [8].

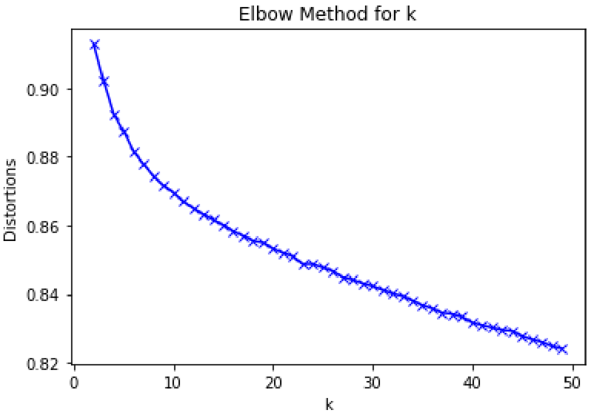


Fig. 6. The elbow method using distortion

Algorithm 1 KMeans Clustering

```
Input: Training set $(x_1, x_2, ..., x_n)$, Number of Clusters $K$
  • Randomly initalize $K$ cluster centroids $\mu_1, \mu_2, ..., \mu_k$
repeat
    for $n = 1$ to $N$ do
        r$_{nk}$ ← argmin$_k$ $\| x_n - \mu_k \|_2$ : Assign data point to closest center
    end for
    for $k = 1$ to $K$ do
        $\mu_k$ ← MEAN$_k$ $(x_n, r_{nk})$ Re-estimate mean of cluster $k$
    end for
until Centroid Position do not change;
  • Return r: Return cluster assignments
```

F. Visualization

The Fig. 7, shows the result of clustering the scholarly articles with 3 clusters. Each cluster represents a group of related articles that can simplify the search for related articles. To get more insight into each cluster and what does the cluster represents, word cloud are used that visualize most important words. The Fig. 8, represents word cloud for cluster 0, i.e., red cluster. From the word cloud an interpretation can be made that this cluster represents article related to influenza, covid, health, outbreak, pandemic etc.

The Fig. 9, represent cluster 1, i.e., green cluster. From the visualization an overview of the cluster can be gained. It shows cluster 1 contains more information about nucleolus, nuclear, defence, immune, health, vaccination etc. Fig. 10, depicts cluster 2, i.e, blue cluster. The cluster is more related to protein, genomic, pulmonary, oxygenase, carbon, monoxide etc. Hence, the clustering and visualization helps to get insight into trends and direction of research that would have been very time consuming and tedious to do manually.

G. Training & Classification

During clustering, the datasets are labeled with the cluster values, so the new data are be considered as labeled data and used for training and classification purpose. The idea behind

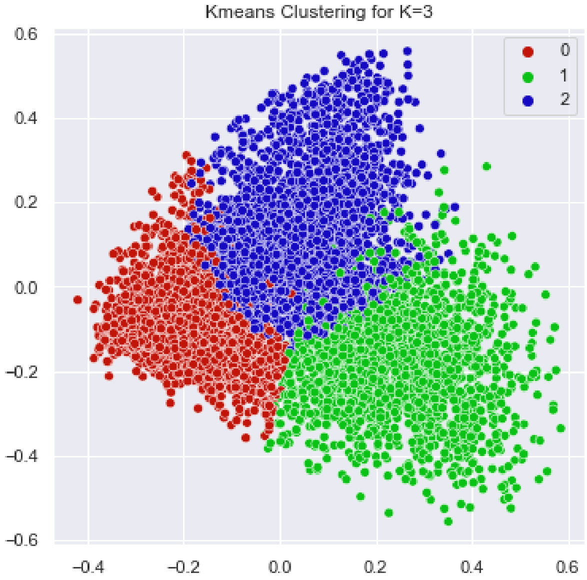


Fig. 7. Kmean Clustering; K=3

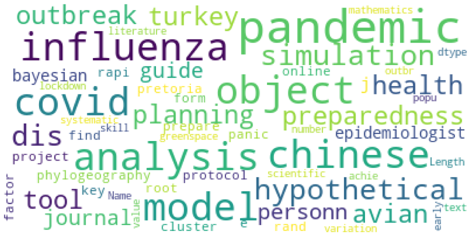


Fig. 8. Cluster 0 (Red)

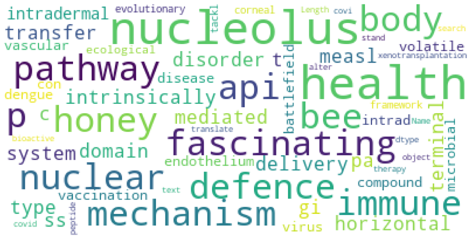


Fig. 9. Cluster 1 (Green)

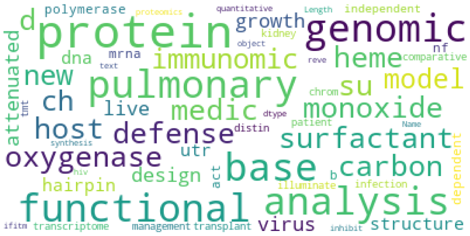


Fig. 10. Cluster 2 (Blue)

this is to evaluate how well the cluster has generalized. Support Vector Machines will be used for training and classification.

The data are splitted into training and testing datasets, where training data account 70% and testing data holds 30% of total datasets.

1) Support Vector Machine (SVM): A Support Vector Machine is a discriminative classifier that can be used for classification as well as for regression problems. This supervised learning technique builds model, assigning each data item to one of the categories and then representing this data points in space, mapped in a way that clearly separates data points of each category. Similarly, each data point of the test set will also be mapped into the same space by predicting their category based on which side of the separation they fall.

The Table 1 shows evaluation measures for the classifier on test datasets.

TABLE I
EVALUATION MEASURES FOR SVM

SVM	Values
Accuracy	0.976
Precision	0.977
Recall	0.971
F-1 Measure	0.974

H. Search with Text Similarity

From above research, clustering and visualization gave an overall insight into scholarly articles information. Researcher who are searching for specific information in a domain need more specific results. As a result, searching articles with text similarity is researched and implemented, so that scientist can find research papers related to particular query or keyword. A technique called Cosine Similarity is used to explore this problem.

The below steps are followed to find top 100 relevant articles about the input query by the user from approximately 10000 scholarly articles.

- A feature vector created using TF-IDF method.
- The query by user is fit into the vector.
- Cosine Similarity between the query and each article are calculated.
- The scholarly articles with highest Cosine Similarity values are presented as result.

1) Cosine Similarity: The angle between two vectors of an inner product space is measured by Cosine Similarity. The cosine of the angle between vectors is concerned with directions and determines if they are in same direction. Two vectors will have cosine similarity value of 1 if they are in the same direction and 0 if they are aligned perpendicular to each other and is used in positive space. It is widely used for document similarity [9].

Mathematically Cosine Similarity starts computing cosine of two vectors derived by Euclidean dot product given as follow [10]:

$$A \cdot B \ = \parallel A \parallel \parallel B \parallel cos\theta$$

Hence, given the dot product of two vectors, cosine similarity can be written as below:

$$Similarity = cos(\theta) = \frac{A \cdot B}{\parallel A \parallel \parallel B \parallel}$$

$$Cosine\ Similarity = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

In general, it computes output from -1 to 1, where 1 indicates total similarity, -1 represents no similarity and 0 is perpendicular. In practice cosine similarity is used in positive space, i.e., from 0 to 1.

I. Evaluation

1) Classification Evaluation: Once classifier are trained on the data provided, the performance of those classifiers need to be evaluated. The following measures are used to evaluate the performance of classifier.

- Accuracy
- Precision
- Recall
- F-1 Score
- *Accuracy:* Accuracy is simply a ratio of correctly predicted observation to the total observations.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made}$$

- *Precision:* Precision is the ratio of observations predicted as positive to all observations that are positive.

$$Precision = \frac{True\ Positive}{True\ Positive + FalsePositive}$$

- *Recall:* Recall is the ratio of observations predicted as positive to all observations that are positive in the actual class.

$$Recall = \frac{True\ Positives}{True\ Positive + FalseNegatives}$$

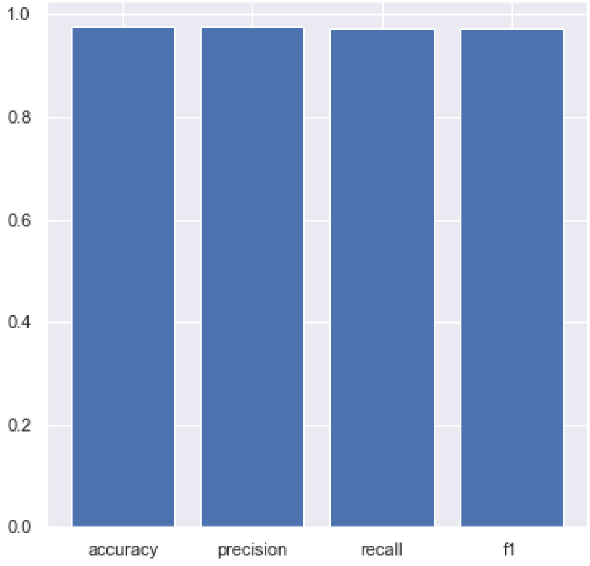


Fig. 11. Classification Evaluation for SVM

- *F-1 Measure:* F-1 Measure is simply the weighed average of Precision and Recall. It is sensitive to both false positive and false negatives.

$$F-1\ Measure = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The Fig. 11, shows the result for Accuracy, Precision, Recall and F-1 Measure in the form of chart to visualize the classification results.

2) Search Evaluation and Test: As mentioned in Cosine Similarity section, the results of search are based on highest cosine similarity values. An example of search with query and its similarity values with results are given below.

$$query = ''influenza\ pandemic\ simulation''$$

The Table 2, shows highest 5 cosine similarity values between query and corresponding article with given index.

TABLE II
COSINE SIMILARITY VALUES

Article Index	Values
5574	0.532
1356	0.506
2146	0.499
1699	0.491
6389	0.488

The Table III, displays top five results from overall datasets. To test how well the outputs are, they can be cross checked with the input query. The title of the top five result shows, they are related about Influenza Pandemic, that matches the search query. These search results are also the best results from the whole datasets as they have highest cosine similarity values.

TABLE III
SEARCH RESULT

Article Index	Result Title
5574	Pandemic Influenza
1356	Global Strategies and Response Measures to the Influenza A (H1N1) Pandemic
2146	Pandemic Influenza: A Comparative Ethical Approach
1699	Seasonal and Pandemic Influenza Surveillance and Disease Severity
6389	Anticipation and response: pandemic influenza in Malawi, 2009

IV. Analysis & Conclusion

The research was focused on clustering, Classification and search of the scholarly articles and visualizing it to get an insight about published research papers. Clustering and visualization gave an overview about what kind of information are the scholarly articles representing. This helps researcher to know the trend and direction of an on going research. Searching relevant articles are important to get specific information, that helps in quick information retrieval and support the research. Mathematical algorithms like KMeans, Principal Component Analysis, Random Forest, Cosine Similarity are used to solve the discussed problems.

V. Future work

As the research was performed in small scale datasets according to computational capacity, the research will be extended in future to cover more datasets. A full fledged software module will be designed that can be easily used by any researchers or users without any technical/computing background.

References

[1] Doanvo, A., Qian, X., Ramjee, D., Piontkivska, H., Desai, A., Majumder, M. (2020). Machine learning maps research needs in covid-19 literature. Patterns, 1(9), 100123.

[2] 2. https://www.semanticscholar.org/cord19

[3] Pereira, J. (Ed.). (2010). Handbook of research on personal autonomy technologies and disability informatics. IGI Global.

[4] Mehryar, M., Rostamizadeh, A. (2012). Ameet Talwalkar Foundations of machine learning.

[5] S. Nikita, ”Understanding the Mathematics behind Principal Component Analysis,” Feb 13, 2020. [Online].Available:https://heartbeat.fritz.ai/understanding-the-mathematics-behind-principal-component-analysis-efd7c9ff0bb3. [Accessed Feb 14, 2021].

[6] G. Ian, B. Yoshua, C. Aaron, “Deep Learning.,” MIT Press, 2016

[7] P. Chris, ”K Means,” 2013. [Online].Available:https://stanford.edu/ cpiech/cs221/handouts/kmeans.html. [Accessed Feb 14, 2021].

[8] D. Dirk-André, ”Advance topics of machine learning ,” Nov 29, 2018. [Online].Available:8. http://www.mathematik.uni-muenchen.de/ deckert/teaching/WS1819/ATML/arman_unsupervised_learning.pdf. [Accessed Feb 13, 2021].

[9] Han, J., Kamber, M., Pei, J. (2011). Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 5(4), 83-124.

[10] ”Cosine Similarity ,” [Online].Available:8. https://deepai.org/machine-learning-glossary-and-terms/cosine-similarity. [Accessed Feb 14, 2021].