Politically-biased news detection using Machine Learning Techniques

Sahiti Cheguru

Student, B.Tech, Dept of CSE,

Gokaraju Rangaraju Institute of Engineering and Technology, Bachupally, Hyderabad, Telanagana

Jawaharlal Nehru of Technological University Hyderabad, Telangana

sahiticheguru2000@gmail.com

Dr.Y. Vijayalata
Professor, Dept of CSE
Gokaraju Rangaraju Institute of Engineering and Technology, Bachupally, Hyderabad, Telanagana

Jawaharlal Nehru of Technological University Hyderabad, Telangana

vijaya@ieee.org

Abstract. News articles have a relatively high trustworthy rate as compared to other platforms when it comes to gathering knowledge and information. However, biased media coverage leads to spreading of wrong information that manipulates people's perceptions. This project joins the powers of neural networks and computational linguistics to detect political bias persisting in news articles. The research progresses with the idea of binary classification of articles by grouping them as biased and non-biased. Deep learning tools such as Bag-of-words, Word embeddings, and Doc2Vec along with Long Short-Term Memory (LSTM) networks are used to identify bias in a training dataset of 600,000 articles and a validation dataset of 150,000 articles.

Keywords: Natural Language Processing, Machine Learning, Media Bias

1 Introduction

Slanted news coverage, can strongly impact the public perception of the reported topics[1]. Given that news articles are the primary source of information for a broad demographic spectrum of readers, it is imperative that unbiased news is more accessible to everyone. Media bias can be in the political domain which involves specifically choosing to highlight some events, parties and leaders. Bias can be detected in these articles by observing the unclear assumptions, loaded language, or lack of proper context[2]. On a different note, evidences such as editors stating that the coverage does not reflect their own ideas but those of the people upon whom the papers depend for revenue. It suggests a connection between consumers' prior beliefs and media firms' slant [3]

This project aims to examine the political bias prevailing in news articles and identify bias indicators by dealing with a large noisy dataset. This noise stems from fact that the machine learning models used tend to learn publisher-specific traits instead of potential individual bias in the articles. The research progresses with the idea of binary classification of articles by grouping them as biased and non-biased. A word vector specific technique called "word embedding aggregation" [4] is used to extract the features. The article text is aggregated into embeddings using Word2Vec embeddings and De Boom et al.'s word embedding aggregated sentence embeddings. The goal in using Doc2Vec is to avoid a neural network learning publisher-specific lexical features, and to instead be able to generalize with the aggregated document vectors. Deep learning tools along with Long Short-Term Memory (LSTM) networks are used to identify bias in a training dataset of 600,000 articles and a validation dataset of 150,000 articles.

2 Corpus

The dataset is drawn from the large corpora published by FactCheck.com and BuzzFeed. The training dataset consists of 600,000 articles a validation dataset of 150,000 articles with half being politically biased.

Test Dataset

These test datasets were obtained via web scraping from news providers focusing on entertainment. The test dataset, here referred to as byarticle dataset includes 638 articles with no publisher overlap with any of the given corpora. The other, like the training and validation sets, was labeled overall by publisher, here referred to as bypublisher-test dataset, with a total of 4000 articles, also including publisher overlap with other datasets.

Preprocessing the data

The noisy dataset was cleaned through tokenization of texts using the Natural Language Toolkit (NLTK). Special characters, double spaces, more than three dots in a row, and any failures in character translation were replaced or removed. Regular expressions were used for some of these tasks, as well as for removing img or html tags and URLs. Another list of phrases was summarily removed from each text: those which were likely byproducts of the articles' retrieval from their websites. These included

"Continue Reading Below...", "Image Source:", "Opens a New Window", and so on.

3 Methodology

After collecting and pre-processing the data, the following NLP tools are used to extract the required insights from it.

Bag-of-Words Unigram Model

The bag-of-words model is used to determine the term frequency of the words of the research interest that indicate bias. The bag-of-words model is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature or training a classifier[5].

First, the plain text is converted to a sequence words using a tokenizer and further reduced into their lemmas. To feed this data to the logistic regression model, each term was transformed into a vector that can be placed in the input layer of a neural network. Scikit-learn is used to create a vocabulary of the most common 4000 words to achieve this in the overall corpus (all datasets), excluding stopwords. Vocabulary words from an exception list was also excluded. This exceptions list was formed by counting all words in the training, validation datasets, and gathering those words which appeared five times more often (relative to the size of the corpus) in one set than in the other.

Feature Assignment

Each term is then characterized with a set of features. In order to capture syntactic relations, dependency and part of speech tags were used in this approach; to keep lexical information involved without assigning inflated importance to publisher-specific words, existing lexicons developed by researchers in the field are used.

Feature	Weight
Adjectives	2.04
Adverbs	4.76
Pronouns	−2.87
Nouns	−0.32
Proper Nouns	−1.04
Strong Subjectivity	3.28
Clues Weak Subjectivity	2.85
Clues Auxiliaries	−2.10
Passives	−2.83
Average Sentence	−0.01
Length Average Word	−0.01
Length Exclamation	0.26
Marks Question Marks	0.40
Multiple Punctuation	0.05
Anger Words	0.06
Fear Words	−2.09
Sadness Words	−0.94
Joy Words	−4.39

Table 1. Features chosen for the second linguistic approach, and their weights according to the trained model.

Part of speech tagging and dependency parsing were both done using spaCy, simultaneously with the tokenization and lemmatization mentioned in the description of the last approach. A Ridge Classifier model from scikit-learn was fit to the vectorized training dataset.

LSTM

Recently, owing to the breakthrough in the field of computational science, deep learning or deep neural network (DNN) methods based on artificial neural networks have received a growing interest both academically and practicality from scientists[6]. Long Short Term Memory networks (or LSTMs) are one of the state-of-the-art DNNs which are capable of learning long-term dependencies.

The text is vectorized in order to be used in the machine learning model. To do so, skip-gram word embeddings on the entire dataset are trained. Embeddings of 50 dimensions were trained using the Python package gensim, for 10 epochs, including words in the vocabulary which appeared in the corpus over five times.

Texts were first transformed into arrays of the shape (100, 50), wherein 100 was the cutoff or maximum text length and 50 was the dimensionality of the word embeddings. Texts shorter than 100 words were padded with zero-vectors to keep the shape consistent to feed into the network. The model consists of a single LSTM layer with 50 units, followed by a dropout wrapper with a keep probability of 0.75 to help prevent overfitting. Next is a standard feedforward neural network output layer. AdamOptimizer was used with a 0.001 learning rate as well as softmax cross-entropy loss for optimization. All LSTM models were trained using Tensorflow for approximately 2 epochs.

Voting System

Three LSTMs as described in the previous section were trained: one each on the training, validation and test datasets. Predictions from each LSTM on each article in each dataset were then collected. The articles which all three LSTM models correctly labeled were pulled into a new dataset labeled agree. This dataset, in total size 162,046 articles with 37% biased and 63% unbiased labels, was what the final model was trained on. Once the new datasets were compiled from the voting system based on the originals, a new LSTM with the same architecture was trained on the combined data.

Doc2Vec

In order to obtain neural network-based document embeddings, the doc2vec algorithm is used that learns features from the corpus and provides a fixed-length feature vector as output. Then, the output is fed into a machine-learning classifier[7]. A doc2vec model is developed using genism to minimize the model's problems in learning publisher's tendencies and not generalizing bias. Doc2vec creates word embeddings as word2vec does, but then can generate a single vector of a certain number of dimensions for any document, by aggregating the word vectors and therefore representing the concept of the whole text. The goal in utilizing this process was that the data fed to the model would carry less information about the specific and potentially publisher-specific syntactic constructions and vocabulary, and more information about the idea of the document as a whole, and biased (or not) relation to its topic.

The doc2vec model is trained over all given datasets, only counting words that appeared in the corpus 5 times or more, 10 times (that is, for 10 epochs). The resulting vector size for the document was set to 50. Total training time took approximately 2 hours and 50 minutes.

The feed forward neural network is built using Keras. It included a single fully-connected layer of 64 nodes followed by sigmoid activation. Cross entropy loss and Adam Optimizer were used for optimization. Training was very fast, at about 76 seconds per epoch, and the model was trained for 5 epochs.

4 Results

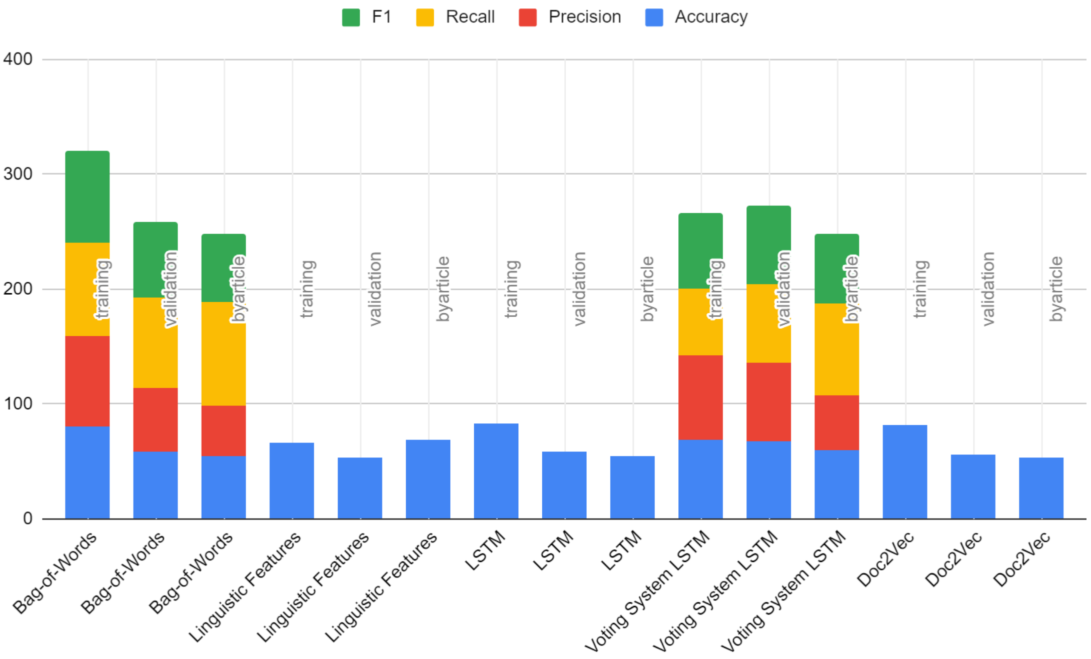

Figure 1. Results on the training datasets.

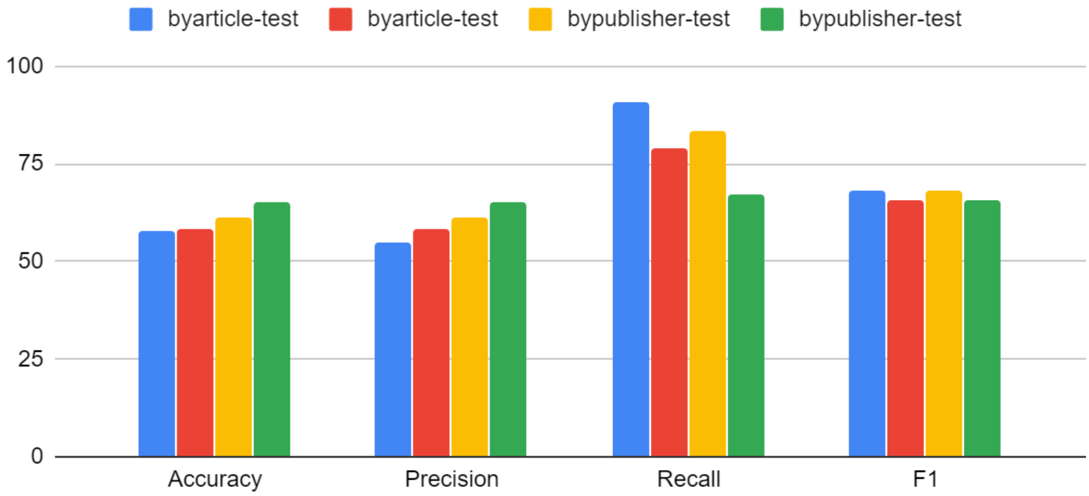

Figure 2. Results on test dataset

Results on available datasets are shown in Figure 1. Generally, the neural network approaches performed better on their training datasets and worse on the validation datasets than the non-deep learning approaches, which seemed to generalize better to other publishers despite higher training error. Interestingly, the results indicate that Linguistic Features approach performed significantly better on the byarticle dataset than the others did.

All results showed higher recall than precision — and except for the case of the Voting System LSTM with the bypublisher-test set, markedly higher. By accuracy, the best result was the Voting System LSTM on the bypublisher-test set.

Overall, the results show a pattern, as briefly mentioned: neural networks seemed to do better on the training dataset and worse on the data belonging to publishers they were not trained on than the bag-of-words and linguistic feature models. This is logical, if the neural networks were able to better pick up on the features that separate publishers from one another, and in doing so generalized worse to other publishers. Essentially: they overfit more strongly to the publishers they were trained on.

Theoretically these articles would all have characteristics common to biased articles of all publishers. When put together, then, the hope was that a model trained on this subset of the dataset would learn those common characteristics and not just the publisher-specific ones. Training the new Voting System LSTM on the pared-down dataset, as expected, reduced training accuracy. Naturally, this could have meant either better generalization to the problem of bias detection or an overfitting to the new dataset instead of just the publishers.

5 Conclusion

This research examines the biased nature of news articles that is commensurate with the bag-of-words model, linguistic features, and also models developed using deep learning. A wide range of approaches were considered to tackle the problem associated with grainy and unstructured datasets. The most onerous task was to train the models characterizing the dataset basing upon the publisher instead of the individual potential bias of the article. Both neural networks with word vectors were used including other machine learning approaches featuring linguistic features, as well as one filtering method for the data using a voting system. Future work involves examining a more rigorous approach to filtering the dataset. This was begun with the Voting System, but could be strengthened. For a better analysis of current results, it would be good to rerun all models for precision, recall and F1 scores, once the computational resources are again available. There are many things that could have gone wrong with this system. Since the three models used for voting did not have very high accuracy on each other's datasets in the first place, the level of noise may not have been reduced at all. Furthermore, the original training dataset was four times as large as the validation dataset. Their subsets after the voting system was applied were equally unbalanced. When combined and used for training the final LSTM, it could have been unbalanced enough that the model learned mostly from the data from the original dataset, and the features from those publishers.

References

1. Hamborg, F., Donnay, K. & Gipp, B. Automated identification of media bias in news articles: an interdisciplinary literature review. *Int J Digit Libr* **20,** 391–415 (2019). https://doi.org/10.1007/s00799-018-0261-y
2. Gangula, R. R. R., Duggenpudi, S. R., & Mamidi, R. (2019, August). Detecting Political Bias in News Articles Using Headline Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 77-84).
3. Gentzkow, M., & Shapiro, J. M. (2006). Media bias and reputation. *Journal of political Economy*, *114*(2), 280-316.
4. De Boom, C., Van Canneyt, S., Demeester, T., & Dhoedt, B. (2016). Representation learning for very short texts using weighted word embedding aggregation. *Pattern Recognition Letters*, *80*, 150-156.
5. McTear, M. F., Callejas, Z., & Griol, D. (2016). *The conversational interface* (Vol. 6, No. 94, p. 102). Cham: Springer.
6. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521, 436.

7. Markov, I., Gómez-Adorno, H., Posadas-Durán, J. P., Sidorov, G., & Gelbukh, A. (2016, October). Author profiling with doc2vec neural network-based document embeddings. In *Mexican International Conference on Artificial Intelligence* (pp. 117-131). Springer, Cham.

JK LAKSHMIPAT UNIVERSITY

ज्ञानम् अमृतम्

JKLU

JK LAKSHMIPAT UNIVERSITY, JAIPUR

INTERNATIONAL CONFERENCE

on

INNOVATION AND SUSTAINABILITY

Certificate of Presentation

This is to certify that **Ms. SAHITI CHEGURU** of **Gokaraju Rangaraju Institute of Engineering and Technology, Telangana** has presented a paper titled **Politically biased news detection using Machine Learning Techniques** in **the International Conference On Innovation & Sustainability** held during **February 5 - 6, 2021**.

Dr. Amit Sinhal
(Conference Convener)

Dr. Lokanath Mishra
(Conference Convener)

Dr. R. L. Raina
(Vice Chancellor, Conference Patron)

JK LAKSHMIPAT UNIVERSITY, JAIPUR

INTERNATIONAL CONFERENCE

on

INNOVATION AND SUSTAINABILITY

Certificate of Presentation

This is to certify that **Ms. SAHITI CHEGURU** of **Gokaraju Rangaraju Institute of Engineering and Technology, Telangana** has presented a paper titled **Group Discussions Analysis and Digression Intervention** in **the International Conference On Innovation & Sustainability** held during **February 5 - 6, 2021**.

Dr. Amit Sinhal
(Conference Convener)

Dr. Lokanath Mishra
(Conference Convener)

Dr. R. L. Raina
(Vice Chancellor, Conference Patron)

Group Discussions Analysis and Digression Intervention

Sahiti Cheguru

Student, B.Tech, Dept of CSE,
Gokaraju Rangaraju Institute of
Engineering and Technology,
Bachupally, Hyderabad, Telanagana
Jawaharlal Nehru of Technological
University Hyderabad, Telangana
sahiticheguru2000@gmail.com

Dr.Y. Vijayalata
Professor, Dept of CSE
Gokaraju Rangaraju Institute of Engineering and
Technology, Bachupally, Hyderabad,
Telanagana.
Jawaharlal Nehru of Technological University
Hyderabad, Telangana
vijaya@ieee.org

Abstract

It is in common knowledge that reading is one of the richest sources of knowledge in this world. Reading empowers you with the light that leads you through the dark. Therefore, we attempt to promote this valuable skill with this study. In this paper, a platform is developed that facilitates the exchange of thoughts and information among students. We have leveraged NLP to develop this application and categorize texts into various categories. Further, various text classification methods are introduced to derive meaningful insights from written communication among students regarding books. We go on to apply the information drawn from text classification to a technology that engages readers through interactive games and discussions, IMapbook. The conversational text acquired through these discussions is further classified into various categories based on the context. Here, we aim to build a classifier that can predict these categories. Our study shows that the fine-tuned BERT, outperforms all the other methods used in this research.

1 Introduction

Nature Language Processing is a burgeoning field that has seen plenty of research breakthroughs recently. It is now broadly studied topic with many successful applications. In this project we touch subfield Text Classification and apply its methods to the data from IMapbook[1], a web-based technology that allows reading material to be intermingled with interactive games and discussions. Some portion of discussions from this platform were manually annotated, each reply was given more categories based on the information in the reply.

Our goal is to take this data and try to build a classifier which would predict these categories. Such classifier could then be used to automate analysis of discussions at this platform, recommend the time for the teacher's intervention. The domain of our problem is short-text classification, which is closely related to social media. Unlike the common text classification problems, where the documents are usually long and written in formal language, it deals with texts of few sentences, written in informal language. The amount of context information carried in the texts is usually very low, thus classification and information retrieval become challenging tasks to perform efficiently Furthermore, the low co-occurrence of words induced by the shortness of the texts of ten results problematic for machine learning algorithms, which rely on word frequency.
With the arise of social media this branch of text Classification became a well-researched problem,

and people tried different approaches to overcome its constraints. Currently, the most widely used vector representations of words (or embeddings), that proved to capture well the semantic information are GloVE [2] and Word2Vec [3].

Although standard machine learning approaches often resulted problematic with short text [4], showed that their model with hand-crafted features, related to user's tweets, efficiently filtered irrelevant tweets from the users, thus suggesting that by adding extra sources of context information increases the performance. Similarly, this concept was also recently shown by Yang et al. [5]. Furthermore, they have also shown that Support Vector Machines performed almost equally well in classification when using word embeddings or TF-IDF, but they were outperformed by deep neural networks.

2 Dataset

The dataset is provided by IMapBook and includes the discussions between students and teachers on the topics of the book they are reading. The dataset includes approximately 3500 Slovene messages, from 9 different schools and on 7 different books, which were also translated to English. Students in each school were divided in "book clubs", where the conversations occurred.

The data was manually annotated with three main tags:

- *Book Relevance*: Whether the content of the message is relevant to the topic of the book discussion.

- *Type*: Whether the message is a question(Q), answer(A)or a statement(S). In original data mixture of these classes also appear (QA and AQ), but because of their low frequency (together they appear only three times in entire dataset), we changed QA occurrences to Q and AQ to A.

- *Category*: Whether the message is a simple chat message (C), related to the book discussion (D), moderating the discussion (M), wondering about users' identities (I), refer- ring to a task, switching it or referring to a particular position in the application (S), or other cases (O).

The *Category* category can be further on split in sub-categories; *chats* may be in the form of greetings (G), related to the book (B), they could be encouraging (E), talk about feelings (F), contain cursing (C) or others (O), *Discussion* messages could be questions (Q), answers (A), answers to users, still related to the discussion topic (AA) or encouraging the discussion (E); *identity* messages can be answers(A),questions (Q) or their combination (QA).

The dataset is suitable for both binary and multi-class classification, whether the target variable is the relevance or the category of the message respectively.

3 Methods

In this segment, we go over the techniques for three different message classification tasks:

1. Book relevance classification(binary)

2. Type of message classification(3-class)

3. Broad category classification(6-class)

Input data to all classifiers are exchanged messages. To provide information about whether users are discussing about relevant topic, each message also has information about the question provided to users before the discussion.

Baseline

As a baseline model we decided to use Majority Classifier. In each task it classifies every instance as the most representative class in training set.

Hand-Crafted Feature Models

The first group of models that we present isbased on a hand-crafted feature set. These features were then used as an input to different classification algorithms, that we list in Section 3.2.2. We describe features extraction in the nextsection.

Features Extraction

The aim of the features was to simply and intuitively capture the relevance to the question, while filtering gibberish and inappropriate messages. Thus, the following set of features was de- signed:

- Tokens in a message.

- Mistakes in a message; this was computed by matching words with the words in a lexicon [6]

- Maximal length of the token in the message.

- Characters in a message.

- Question marks in a message.

- Exclamation points in a message.

- Commas in a message.

- Periods in a message.

- Capital letters in a message.

- Capital letters within the interior of the words in a message.

- Peculiar characters in a message.

- Numbers with in the interior of the words in a message.

- *Levenshtein distance:* Number of all pairs of words from the question and the message, whose Levenshtein distance is less than half the length of the longer of the two words.

- Interrogative words in amessage.

- "kdo" in a message.

In the case of *Levenshtein distance* feature, the messages were initially tokenized and stop-words [7]were removed, while for other cases regular expressions were used to extract the features.

All features were designed while looking at the data, having some sense in how the feature could increase the classification success. For instance, many messages had "kdo" word in it, asking for identity of somebody. Those messages have the same class. But nevertheless, we observed only small portion of the data, so that chosen features would not be overfitted.

Classification Algorithms

We decided to feed the features to four different classification algorithms to see how they perform. We chose a Naive Bayesian (NB), random forest (RF), support vector machine (SVM) and a logistic regression (LR) classifiers. We used the implementations from scikit-learn library [8].

When selecting the parameters we observed train and test accuracy and paid close attention to detecting over fitting. For NB we left the default parameters. For the SVM we used the RBF kernel and set the parameter *gamma* to "auto" and *C* to 5, while for the LR we decided to use "lbfgs" optimizer with maximum 1000 iterations. In the case of LR the input data was standardized to ensure equal class importance. For the RF we set the number of estimators to 150, while *min_samples_leaf* to 3 and *min_samples_split* to 10. This way we managed to reduce the over fitting to the training data. We kept the same parameters for all the tasks.

ELMo Embeddings

We handcrafted features by looking at the messages and observed what could potentially discriminate different types of messages. For the next experiment we wanted to know, how good features can we extract automatically, so that such human interaction and understanding of messages wouldn't be necessary.

ELMo [9] is model for creating contextual embeddings. We have chosen it as it can also be used to embed entire message. Firstly, we put discussion topic into it, and then message, so that message's embedding also contains information about the relevance to the topic.

We have used pre-trained ELMo model for Slovene language [10].

For classification we tried all models discussed in 3.2.2 and also KNN [11] with cosine distance, as it is natural distance to use in ELMo embeddings. Random Forest classifier ended up having the highest performance.

Fixing Typos inMessages

Messages in the input data contain a lot of words that have typos in them and are not part of the Slovene lexicon [6]. Also, a lot of mistakes come from users deliberately leaving out carrot (e.g.'s' instead of 'š'). That is why we decided to write an algorithm for correcting typos that are away from the correct word for at most Levenshtein distance of two. We also calculated probabilities of the words and removed words with probability less than10^{-8}.

BERT Fine-Tuning

Another approach we propose is using a pre-trained BERT [12], by fine-tuning it for our classification tasks. We avoided customizing BERT models, because they require notorious amount of data which was not available. We trained our BERT model for Slovenian, Croatian and English languages for sequence classification for three epochs on training data that consisted of 80% of our dataset, while the remaining 20% was left for testing. Out of these 80%, 15% were used for validation. We trained one model for each task, for both Slovenian and English- translated messages.

4 Evaluation

We evaluated the models using F1 evaluation metric. At multi-class problems we used weighting over different classes to compute it.

Because of the complexity of our models, we opted for two different evaluation techniques: on models that are not so computationally expensive tofit, 5-fold cross validation was applied, where our performance estimator was the average result of the five test sets. This technique also points out the variance of our estimator, hence quantifying to some extend the uncertainty of the performances of our models. The second evaluation is a simple hold-out evaluation, where we split train and test sets at 80%, thus losing information about the variability of the performance of our predictor.

5 Results and Evaluation

BaselineModels

Scores for computationally less expensive models (Majority Classifier and different models with handcrafted features) are shown in Figure 1. We notice that all feature-based classifiers outperform the Majority Classifier. Furthermore, as expecting, the classification accuracy drops within creasing number of target classes. The best performing classification algorithm on this dataset is Random Forest, which outperformed the others in both "Category Broad" and "Type" classification tasks. Its performance on the “Book Relevance “task was also on average higher than the rest. However SVM and LR obtained comparable results.

Initially, RF yielded very high performance on the training set, reaching 95% accuracy. However, the performance on the test set was lower, showing signs of over fitting. Thus, with a more careful selection of the parameters, we dropped the training accuracy for about 10% and reached the current test performance.

Features Importance

RF is often used as a feature’s selection tool, as it ranks the importance of the features. The features occupying the upper section of the tree are majorly decisive in predicting the output. The inputs taken for this purpose can be used analyze the most important features or gauge their relative importance. Figure 2 spotlights the vitality of every feature in the decision process of the Random Forest model.

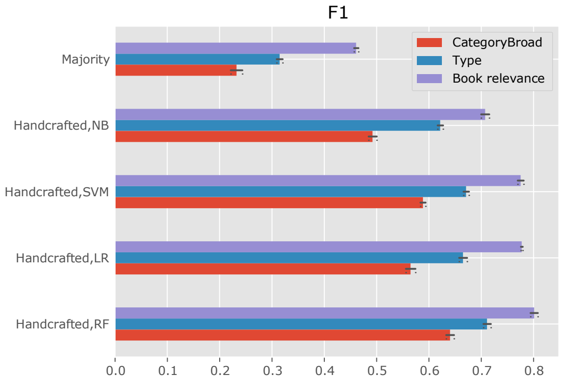

Figure1:**F1 scores** Scores of baseline models on three different classification tasks described in Section2.

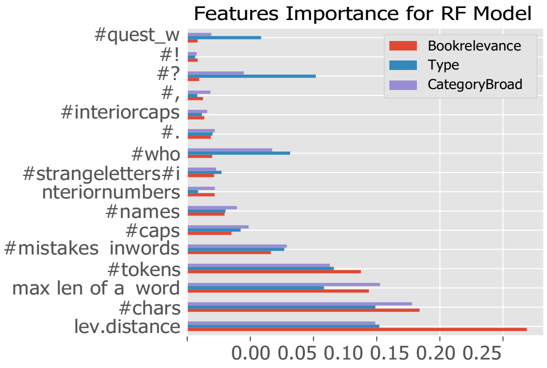

Figure 2: **Features Importance.** *Lev. distance* between answer and question, general length of message, and number of mistakes show as important features.

As we notice, each classification task focuses on different features. However there are some common ones that are discriminatory for all three tasks, i.e. last five in the plot. As expected, *Lev. Distance* works particularly well on the "Book relevance" problem since it performs a naïve kind of matching of the text messages with the questions. However, it results also as the most discriminatory feature for "Type" classification and third for "Category Broad" classification.

It is not surprising that some features are particularly relevant to some classification tasks, since they were designed for that purpose. It is also known that good features increase performance. Here we showed that some features are particularly suitable for some specific tasks, while others behave well over different classification problems. One future improvement that could be done is trying to define some other features that would boost the performance, removing the irrelevant ones.

Deep Models

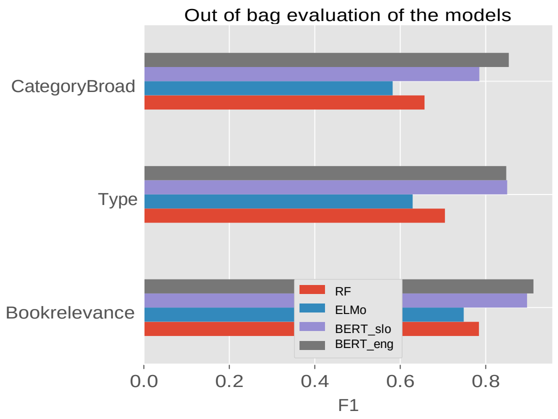

Figure 3: **Hold-out performance evaluation.** Comparing performances on the test set of BERT, Handcrafted Features Model and ELMo model.

F1 scores from hold-out evaluation are shown in the figure above and in this table below.

	Relevance	Type	Category.
Handcrafted	0.78	0.70	0.66
ELMo	0.75	0.63	0.58
BERT	0.90	**0.85**	0.78
BERT (Eng)	**0.91**	0.85	**0.85**

Here BERT (Eng) is BERT trained on English translations. Note that these translations were made by human and wouldn't be present in unseen data.

ELMo

We can see that ELMo has worse performance than baseline model with handcrafted features. But it is important to note here, that handcrafted features may be over fitted to the given data. If model was applied to discussions from older children, same features may perform worse. In the other hand, ELMo features are generated automatically and may generalize better.

BERT

When analyzing performance of the BERT models, we can clearly see15-20% increase in performance compared to model with hand crafted features. BERT model that uses English translations is even more successful, especially in the classification of the category, where we can observe nearly 29% increases in performance. This clearly demonstrates dominance of BERT models.

We would like to mention, that here we did not measure uncertainty of the scores. But scores are still comparable, as we evaluated models on the same test set.

Analyzing Predictions

A lot of messages are asking for identity of somebody, and such messages were mostly successfully classified by all models. Lots of messages contain a lot of gibberish and are as such distinguishable from other messages. Harder to predict are messages that are short and contain only few words. Models performed worse also on messages with a lot of unidentified mistakes in words.

6 Conclusion

Reading opens the gates to knowledge and wisdom like nothing else in this world. This paper progresses with this idea while drawing the benefits of Natural Language Processing. We experiment through text classification methods to understand the degree of accuracy to which they can automatically assign relevant categories to pieces of text. As a baseline model, we decided to use Majority Classifier. We chose to feed features to Naïve Bayesian (NB), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) classifiers for this purpose. Further, we worked with Elmo Embeddings where Random Forest delivered the highest returns. We also fine-tuned the end-to-end BERT neural network, yielding a significant increase in performance. Future work for this research involves the recognition of messages that are direct replies to a particular message. This would improve classification with additional context that will make the categorization more meaningful.

References

[1] Grandon Gill and Glenn Gordon Smith. 2013. Imapbook: Engaging young readers with games. *Journal of Information Technology Ed- ucation: Discussion Cases*, 2(1).

[2] Jeffrey Pennington, Richard Socher, and Chris toper Manning. 2014. Glove: Globalvectors for word representation. volume 14, pages 1532–1543.

[3] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

[4] Bharath Sriram, Dave Fuhry, Engin Demir,Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. *Proceedings ofthe 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages841–842.

[5] Xiao Yang, Craig Macdonald, and IadhOunis. 2018. Using word embeddings in Twitter elec-tion classification. *Information Retrieval Jour- nal*, 21(2-3):183–207.

[6] KajaDobrovoljc, Simon Krek, Peter Holozan, TomažErjavec, Miro Romih, ŠpelaArharHoldt,JakaČibej,LukaKrsnik,and Marko Robnik-Šikonja. 2019. Morphological lexicon sloleks 2.0. Slovenian language resource repository CLARIN.SI.

[7] JožeBučar.2017.Automaticallysentimentan-notated Slovenian news corpus Auto Senti News 1.0. Slovenian language resource repository CLARIN.SI.

[8] F. Pedregosa, G. aroquaux, A. Gramfort, V. Michel, B.Thirion, O. Grisel, M.Blondel, P.Prettenhofer, R. Weiss, V.Dubourg, J. Van- derplas,A.Passos,D.Cournapeau,M.Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit- learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[9] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

[10] MatejUlčar.2019.ELMoembeddingsmodelsforseven languages. Slovenian language resource repository CLARIN.SI.

[11] Keinosuke Fukunaga and Patrenahalli M. Narendra. 1975. A branch and bound algorithm for computing k-nearest neighbors. *IEEE transactions on computers*, 100(7):750–753.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre- training of deep bidirectional transformers for language understanding. *arXivpreprint arXiv:1810.04805*.

COVID-19 Forecasting Using Deep Learning Models

Abstract:
COVID-19, responsible of infecting billions of people and economy across the globe, requires detailed study of the trend it follows to develop adequate short-term prediction models for forecasting the number of future cases. In this perspective, it is possible to develop strategic planning in the public health system to avoid deaths as well as managing patients. In this paper, forecast models comprising various artificial intelligence approaches such as support vector regression (SVR), long shot term memory (LSTM), bidirectional long short term memory (Bi-LSTM) are assessed for time series prediction of confirmed cases, deaths and recoveries in ten major countries affected due to COVID-19. The paper also reviewed deep learning model to forecast the range of increase in COVID-19 infected cases in future days to present a novel method to compute equidimensional representations of multivariate time series and multivariate spatial time series data. The paper enables the researchers to consider large number of heterogeneous features, such as census data, intra-county mobility, inter-county mobility, social distancing data, past growth of infection, among others, and learn complex interactions between these features. To fast-track further development and experimentation, the analysed code could be used to implement the AI in an efficient way. The paper discusses existing theories and researches that provides a better understanding of the spread pattern recognition which will help to tackle any future pandemic of similar intensity. We encourage others to further develop a novel modelling paradigm for infectious disease based on GNNs and high resolution mobility data.

Keywords. **COVID 19, Artificial intelligence, Deep learning model, Time series data, Prediction model**

1. INTRODUCTION

Coronaviruses earn their name from the characteristic crown-like viral particles (virions) that dot their surface. This family of viruses infects a wide range of vertebrates, most notably mammals and birds, and are considered to be a major cause of viral respiratory infections worldwide.[1] With the recent detection of the 2019 novel coronavirus (COVID-19), there are now a total of 7 coronaviruses known to infect humans. Prior to the global outbreak of SARS-CoV in 2003, HCoV-229E and HCoV-OC43 were the only coronaviruses known to infect humans. Following the SARS outbreak, 5 additional coronaviruses have been discovered in humans, most recently the [2]novel coronavirus COVID-19, believed to have originated in Wuhan, Hubei Province, China. COVID-19 effect has highly noticeable in dense areas with elderly people and people with comorbities. It is considered as multidisciplinary issue for the medical specialists, pharmaceutical industry, local government/health authorities and epidemiological experts. This study is mainly focused on the review of forecasting and prediction of COVID-19 using various deep learning algorithms. A big challenge ahs been witnessed in various science domains globally to restrict the increasing COVID spread trends. Various modelling, forecasting and analysis approaches are established to handle and insight this current pandemic. The evolution of confirmed COVID cases forecasting have been estimated by multiple mathematical models[3, 4].

This study is mainly focused on the review of forecasting and prediction of COVID-19 using various deep learning algorithms. A big challenge ahs been witnessed in various science domains globally to restrict the increasing COVID spread trends. Various modelling, forecasting and analysis approaches are established to handle and insight this current pandemic. The evolution of confirmed COVID cases forecasting have been estimated by multiple mathematical models. This study is aimed at deep learning models and a comparative study is made for forecasting COVID-19 cases. The deep learning models such as Long short term memory LSTM, Bidirectional LSTM, Gated Recurrent unit- GRU and Recurrent neural network- RNN have been analysed. These models possess various advantages like distribution free learning models, managing temporal dependencies in time series data and non linear features modelling of flexibility. Various datasets have been utilized in various studies like John Hopkins dataset from starting to now COVID 19 status. The comparative study and challenges are exhibited in this study.

The major contribution of this study involves,

- To review the various deep learning models related with COVID-19 forecasting and time series prediction globally.
- To analyse the LSTM, Bi-LSTM and GRU techniques applied in various medical images related with COVID-19 cases.
- To made a comparative study for the discussion related COVID-19 prediction and forecasting.

The following section 2 describes the deep learning models against covid-19 and its applications, section 3 describes population attributes of COVID-19, followed by section 4 describes the various deep learning models and the involved COVID-19 dataset. Finally, the conclusion is presented in section 4.

2. DEEP LEARNING AGAINST COVID-19

With the regular increase in the newly acquired and suspected COVID 19 cases, diagnosis of the disease is becoming a growing issue in most of main hospitals because of the inadequate supply of detection systems in the corresponding epidemic area. Radiography and computed tomography hence originated as the integrative players in the pre-detection and diagnosis of COVID 19. But due to the aforementioned overwhelming patients, there occurs false positive rates leading to urgent requirement of computer automated diagnosis like deep learning that precisely confirm patients, screens them thereby conducting viral surveillance. The following studies developed deep learning process on the basis of CT diagnosis for the detection of COVID 19 patients that could able to automatically retrieve the radiographic characteristics of the novel virus, particularly the GGO (ground glass opacity) from the radiographic images.

[5] developed a DL framework for the automatic quantification and segmentation of the quantification of the infectious areas and the whole lung from the corresponding chest scans. The paper employed VB-Net NN (neural network) for the segmentation of COVID 19 infection areas in CT images. This setup has been trained with the utilization of two hundred and forty-nine COVID patients followed by the validation of three hundred patients. For accelerating the manual description of CT images to train the features, a HITL (human in loop) has been adapted for assisting the physician for refining automatic annotation in every case. The assessment of the DL based performance system in accordance with Dice similarity coefficient, percentage of infection in between the manual and automatic segmentation outcomes on the validated images.

[6] provided a fully automated and rapid diagnosis of COVID 19 by adopting deep learning. The experimental assessment on 6524 X rays of various institutions described the efficiency of the suggested method with the average detection time of 2.5 seconds as well as with average accuracy of 0.97.[7] formulated the task of classifying viral pneumonia from the healthy controls and non-viral pneumonia into anomaly detection problems. Hence the study suggested a CAAD model that consisted of shared feature extraction, prediction module and detection module. The main benefit of the suggested method over the binary classification is the

preventing individual class explicitly followed by the complete treatment. This suggested model possess greater efficiency of AUC 84% and sensitivity of 72%.

[8]evaluated the longitudinal modifications of pneumonia in various COVID 19 clinical types at the baseline and following up imaging with the use of quantitative image parameter that has been automatically developed by deep learning system from chest X rays. The major findings of the study are lung opacity burden, entire lung and per lobe comparison. This system could able to assess quantitatively the percentage of lung opacification and the recent vision required for the radiologist supervision. The study yielded 8.7% of the cases for insufficient segmentation that ensure precise quantification.

2.1. Medical image processing

Medical image processing is a complex method and understanding of these process is a main cause in the patients who does not respond to the CRT. The study [9] demonstrated the voltage dependent right ventricle capture by the misplaced right atrial lead. The study suggested that device interrogations with the 12 lead ECG and succeeding multimodality imaging must be regarded in accordance to the premature diagnosis of non-responder.

[10] The study aimed to offer burnout medical professions an opportunity by intelligent DL classification methods. The study detected an appropriate CNN model by an initial comparative analysis of various CNN framework. The study then optimized the selected VGG 19 model for image modelling for depicting that the model might be utilized for high demand and challenging datasets. The paper then highlighted the limitations in using the publicly available datasets for the development of useful DL models and the process of creating adverse impact on training the complex system. The study also suggested an image pre-processing stage for creating a trustworthy dataset in order to develop and test the DL models. This robust method has been aimed to decrease the unwanted noise from the images thereby DL models could focus on the identifying diseases with peculiar features from the extraction. The results represented that the US images offer extraordinary detection rate when compared with the CT and X-ray scans. These experimental outcomes signified that with the presence of limited data, many deep networks suffer for training effectively and provide low consistency when compared with the three used image models. The selected model has been then widely tuned with the corresponding parameters and made to perform the COVID 19 detection over pneumonia or normal lungs for all the three lung models with the accuracy of 84% of CT, 100% of US and 86% of XRay.

Advanced AI methods[11] like deep learning depicted high efficiency in the detection of patterns like the diseased tissue. This study examined the efficiency of VGG 16 base DL model for the detection of COVID 19 and pneumonia with an employment of torso radiographs. The results depicted that high level of sensitivity in the detection of COVID 19 associated with high level of specificity represented that this model could effectively be sued as the screening test. ROC and AUC Curves are higher than 0.9 for all the considered classes.

2.2. Forecasting COVID-19 series

[12]employed six machine learning methods such as CUBIST, RIDGE, RF, SVR and stack ensemble learning and ARIMA model for the cumulative confirmation of COVID 19 in ten Brazil states in accordance with the incidence. The study evaluated the stability of the efficiency and out of sample errors by blox plots. The study failed to adopt DL approach in combination with the ensemble learning. The study did not attempt couplas function for dealing data augmentation. Also the study multi objective hyper parameter tuning hyper parameters for adapting forecasting of the upcoming cases of COVID 19.

[13]focused on two main problems which are as follows: One which generate real time forecast of the upcoming COVID 19 case for several countries and next is the assessment of risk of novel COVID 19 for few more affected countries by a determination of several significant demographic features of the countries and its disease characteristics. For resolving the initial problem, the study presented a hybridised approach on the basis of autoregressive integrated moving model and a wave-let based forecast model for generating short term forecast to determine future predictions of the outbreak. This study might be useful for the efficient allocation of the medical professionals and also it acts as an early warning framework for the government policy makers. Next issue could be solved by the application of optimal regression of tree algorithm in order to determine the important causative variables which considerably affect the fatal rates for various countries. This analysis would necessarily offer deep insight for understanding the early risk of assessing 50 highly affected countries.

2.3. Deep learning and IoT

Because of the global pandemic, there is an emergency requirement for the utilization of technology to their optimum potential. IoT is considered as the one of the recent methods with great capability in performing against the COVID 19 outbreak. IoT comprised limited network where IoT devices sense the surrounding environment and sends useful data on internet.[14]examined the present status of IoT applications in relation to novel virus for the identification and deployment of their operational challenges and suggested the possible outcomes for further pandemic situation. Apart from that the study performed statistical analysis for the implementation of IoT where the external and internal factors are being discussed.

[15]tested several number of COVID 19 diagnosis methods that depend on deep learning algorithms with the corresponding instances. The test results of the study depicted that DL models did not considered defensive frameworks against adverse probabilities that remain vulnerable to the corresponding attacks. At last the study presented in detail regarding the implementation of the attack model of the prevailing COVID 19 diagnostic applications. The study hoped that this process will generate awareness of the adversarial attacks thereby encouraging other for safeguarding DL methods from the attack of the healthcare system.

[16]investigated the insight of DL tool application from the diverse view for empowering IoT applications in 4 major domains comprising smart home, smart health care, smart industry and smart transportation. The main thrust has to be seamlessly coincide with the two divisions of DL and IoT that resulted in an expensive range of new framework in application of IoT like health monitoring, indoor localization, disease analysis, intelligent control, traffic monitoring, home robotics, autonomous driving, traffic prediction, , and manufacture inspection. The study discussed the problems, future research and challenges that use DL and for the motivation regarding further improvement in the promising area.

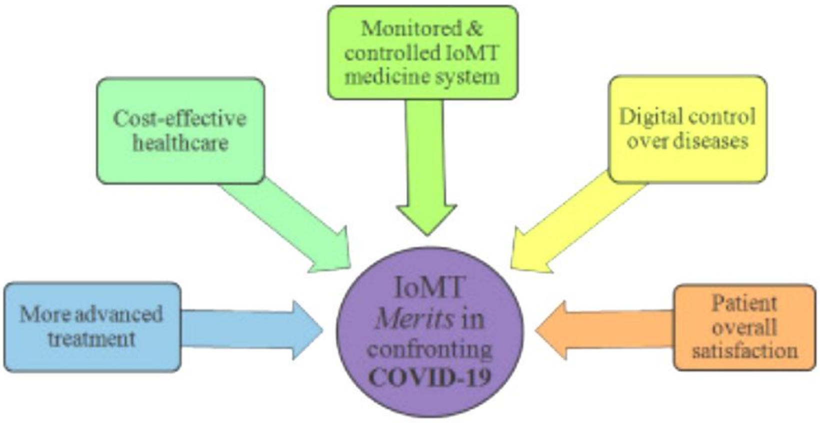
Monitored & controlled IoMT medicine system
Cost-effective healthcare
Digital control over diseases
More advanced treatment
IoMT *Merits* in confronting COVID-19
Patient overall satisfaction

Fig.1. IoT merits towards COVID 19

2.4. *NLP and deep learning tools*

[17]utilized an automated extraction of the coronavirus discussion from the social media and NLP method on the basis of topic modelling for uncovering several issues in accordance with the viral symptoms from the public opinion. Further the study also investigated the usage of LSTM RNN for the sentiment classification of COVID 19 comments. The findings of the present study focussed on the significance of the decision making of COVID 19 issues.

[18]detected and analysed sentiment emotions and polarity that has been described during the beginning of initial stage of the pandemic lockdown period with an employment of NLP and DL techniques on Twitter posts. LSTM models utilised for he estimation of emotions and sentiment polarity from the tweets extracted were trained to obtain existing accuracy on sentiment 140 dataset. This use of emotions depicted a novel and unique method of estimating and validating the supervised learning models on the tweets extracted from Twitter.

2.5. *Deep learning in computational biology and medicine*

Advances in technology in imaging and genomics led to the explosion of cellular and molecular profiling of the data from huge number of samples. This tremendous rise of the biological data acquisition and dimension rate is a complex and conventional analytical strategy. The modern ML methods like deep learning promise to handle huge datasets for the determination of the hidden structure within them thereby making precise predictions. The review discussed the application of novel breeds and approaches in cellular imaging and regulatory genomics. The study provided a background of the summary of deep learning and provided certain tips for the practical usage with possible pitfalls and challenges for guiding the computational biologists in the utilization of this methodologies[19].The study[20] briefly introduced the following manuscripts and discussed their overall contribution in the advancement of science and technology: transcriptomic, cancer informatics, visualization and tools, computational algorithms, microbiome research and deep learning.

Table 1: Comparative study of the prevailing literatures

S.NO	AUTHOR	DESCRIPTION AND METHODOLOGY	COMMENTS ON THE RESULTS
1.	[21]	The study introduced a novel DL framework (COVIDX-NET) for assisting the radiologists in the automatic detection of coronavirus presence in X-Ray images. This suggested framework comprised 7 various architectures of deep CNN like VGG 19 and the Google MobileNet (second version)	The study described the useful implementation of DL models for the classification of COVID 19 in the COVIDNet processed X Ray images and supported further research in deep learning for diagnosing COVID 19 with high accuracy.
2.	[22]	The study utilized DL model for the automated identification of anomalies in chest CT of COVID 19 patients and compared the quantitative estimation with the radiological residents. A deep learning algorithm comprised of detection of lessions, segmentation and location has been trained and validated in a 14 435 patients with definite pathogenic inclusion.	The suggested algorithm depicted excellent efficiency in a detection of COVID 19 pneumonia on the chest CT when compared with the existing radiologists.
3.	[23]	The issue of automatic classification of pulmonary diseases, comprising the	The results suggested that training CNN from

		recently emerged COVID-19, from X-ray images has been focussed in the study. In specific the existing CNNN known as Mobile net has been employed and trained from the scratch for the investigation of significance of the features extracted for the classification task.	scratch revealed vital biomarkers but not constrained to the COVID-19 disease, whereas the top classification accuracy suggested further analysis of the X-ray imaging potential.
	[24]	The paper assessed the usefulness of the (ARIMA) model in the prediction of the dynamics of Covid-19 incidence at various stages of the epidemic, from initial growth phase, to the maximum daily incidence, until the phase of the epidemic's extinction	The study recommended ARIMA model for forecasting COVID 19 for countermeasures.
	[25]	The study developed prototype of a decentralized IoT based biometric face detection framework for cities under lockdown during COVID-19 pandemic. The study built a deep learning framework of multi-task cascading for the detection of the face.	The study proved that it has an edge over cloud computing architecture.
	[26]	The study built an automated tool known as COVID 19 sign sym that could extract symptoms with their eight	The information extracted is also been mapped to the standardised clinical

		factors (severity, body location, condition, uncertainty, temporal expression, negation subject, and course) from the clinical text.	concept in the general OHDSI model. The evaluations of he notes followed by the medical sayings describe promising outcomes.
	[19]	Explored the possibility of Zakat and Qardh-Al-Hasan as a financial method to handle the adverse impact of Corona virus on poor and SMEs. It resolved by proposing an Artificial Intelligence and NLP based Islamic FinTech Model integrated with Qardh-and Al-Hasan Zakat	The study revealed that Islamic finance has immense potential to overcome any kind of pandemic like COVID 19
	[27]	The study signified the difference and similarity in extensively utilized models in deep learning studies, by discussing their basic structures, and reviewing diverse disadvantages and applications	The study anticipated the work can serve as a meaningful perspective for future development of the suggested algorithm in computational medicine.
	[28]	The paper investigated the networks of non-work related activities in migrant workers to intimate the improvement of lockdown exit techniques and upcoming pandemic preparedness	The study recommended social and geospatial distance followed by avoiding mass gathering and it also encouraged the

		It was conducted with 509 migrant workers over the nation, and it evaluated dormitory attributes, mental health status and social ties, physical and COVID-19-related variables and mobility patterns with the use of grid-based network questionnaire.	welfare of migrant workers.
	[29]	The study assisted the policy makers in taking required decisions in order to stop the pandemic spread, precise forecasting of the propagation of the disease is the paramount significance. The suggested method initially groups the countries possessing same socioeconomic and demographic details as well the health sector indicators with the use of k means algorithm	The method obtained high accuracy in forecasting the daily cumulative viral cases.
	[30]	The study might be used to differentiate several respiratory patterns and the suggested device could be readily employed to the practical utilization.	The suggested deep learning possesses the vital potential to be extended to large scale applications like sleeping scenario, public places and office environment.

3. POPULATION ATTRIBUTES – COVID-19

This study emphasized on the impact of COVID-19 for the migrant workers who are affected immensely. The geographical assessment analysis has been focused and the key facts to control this epidemic has stated. the population attributes are shown in fig.2. The structural barriers have been addressed. The intervention focal points recognized by built environments and social networks. The risk roles of migrant workers in Singapore is thus identified by network’s protective roles [28]. The public health and world economy highly affected due to the COVID-19 pandemic. This kind of issues have been controlled by non-pharmaceutical interventions and this stud utilized the Susceptible Exposed Infected Recovered-SEIR for pandemic dynamics simulation utilizing the society following government, people and business. With respect to social co-operation, the higher realistic implementation related with various social interventions followed. Further COVID ABS model has developed by Python language. By modifying the input parameters this developed model can be extended to other population/societies. For health and government authorities this model s very helpful [31]. In Israel 271 localities have been assessed during the outbreak of 3 months in which 90 percent of population is urban. Higher infection rates seen in political minority groups. On the urban political attributes, the density’s influence and significant impact has highly recorded. Among the environmental degradation and urban sprawl the contagious disease spread leads to new tensions in cities observed from assessment [32]. For population criteria the weight assign is performed by potential approaches which describes the COVID-19 spatial distribution and however the temporal variation has not considered as drawback. The uniform infection rates have not recorded the COVID-19 transmission dynamics. The standard model SEIR has used and it not measured the temporal variation. This study focused in the Brazilian health care system to take an account for the infected patients count. If the control strategies have been affected the infection rate of long term due to the unclear findings [33]. This study major aim shows the infection or death rates have not predicted before or disease evolution. At-risk population have highly focused and the non-hotspot districts characteristics have been analysed. however, from the below graphical fig.1. it shows that the districts with no infections are mostly the rural areas. For denser areas in India, the COVID-19 present burden is higher which is usually the urban areas. For this critical illness the older people shows the larger share of risks [34].

This study developed the contact tracing app in Netherland and the dynamics are not considered. The potential uptake alone predicted from the contact tracing app. For this app promotion the government and local health authorities put lot of effort. Personal data sharing has increasing due to this app and the respondents may changes in future as the disease risks eased [35]. This study utilized the long term climatic records of population density (PD), air temperature (T), specific humidity (SH), rainfall (R), wind speed (WS) with topographic altitude (E), actual evapotranspiration (AET) and solar radiation (SR) at the regional level for the spatial relation association with COVID-19 infection count. With number of infected cases in India of 36 provinces the bivariate analysis shows failure in identifying the important relation. The higher importance has been identified by the partial least square technique. After the analysis of various parameters, the present study focused in India shows the COVID-19 infection are more prone to the hot and dry regions with below altitude [36]. The health population is highly infected by the asymptomatic, symptomatic and pre-symptomatic persons. Another study depicted that the population of asymptomatic patients are higher compared with symptomatic patients. This study has conducted in India and the improved SIERD model utilized to predict both kind of infectious persons. The asymptomatic infected population dynamics evaluated and this study suggested by making these persons into quarantined the number of symptomatic persons also reduced [37].

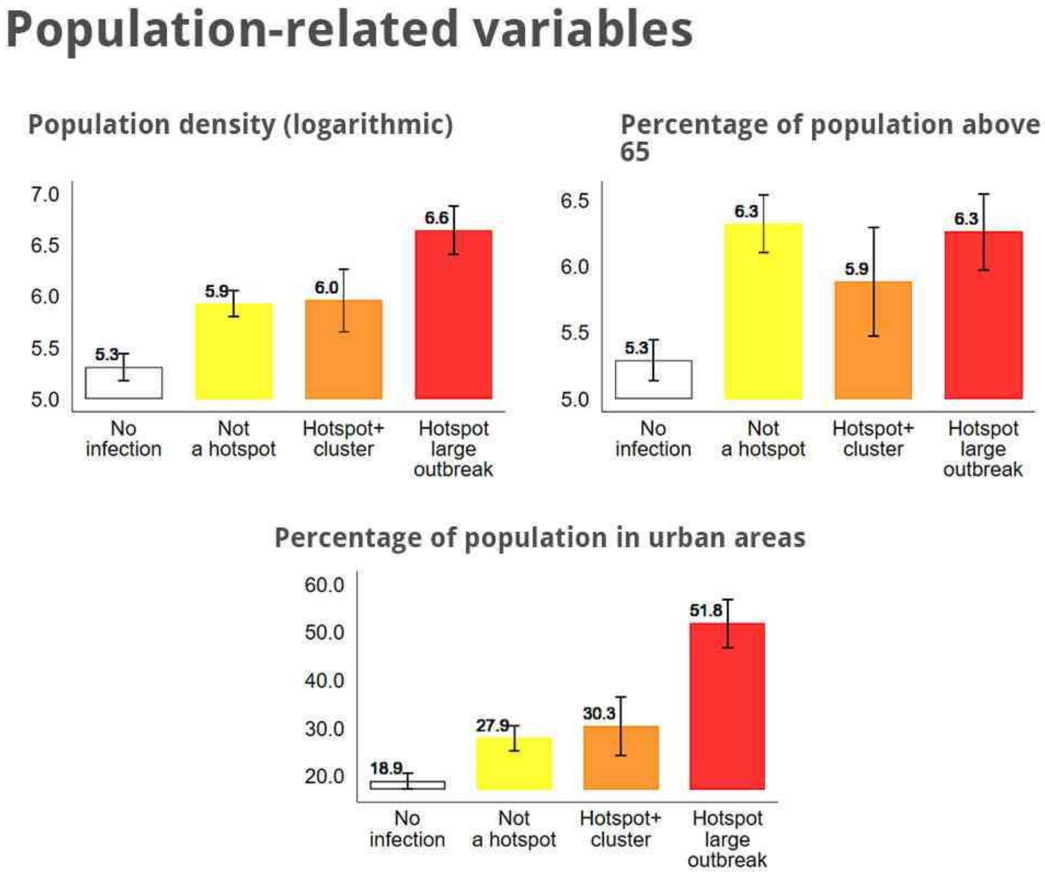

Fig.2. Population variables in India (Source: Scroll.in/National Family health survey data)

4. VARIOUS DEEP LEARNING MODEL

Promising results obtained from the highly challenging state of art methods related with deep learning. The features interpretation and minimal neural architecture is the challenging one. Various deep learning models like CNN, R-CNN, adversarial models, generative and attention based models have been analysed in this study. For image segmentation, various analysis and strong research directions have been estimated [38]. For COVID-19 infection prediction the deep learning models found to be the most appropriate one, according to this study. The personal risk scores from lab assessment assigned for the scarce healthcare resources. From this study, the healthcare resource prioritization improved and patient care has been further informed [39]. For predicting the COVID-19 cases of positive this research proposed the deep learning models. State wise comparison has been made based on mild, moderate and severe in COVID cases. On 32 states, the bi directional LSTM, deep LSTM and convolutional LSTM have been used for an efficient prediction in which maximum accuracy and absolute error has chosen. Bidirectional LSTM shows better results. For the short term prediction for eg. 1 to 3 days BI-LSTM shows better results and it is available publicly. For handles the medical infrastructure these predictions are very helpful for the health authorities. This proposed model can be applicable to all nations worldwide [40]. Based on chest X-ray images, three deep CNN approaches have been utilized for COVID-19 detection. With various kernel functions, deep CNN with SVM classifier has been associated. The results depicted as, this study outperformed the local existing approaches. Compared with deep feature extraction fine-tuning and end to end training needs higher time. Cubic kernel function shows superior performance. Usually ResNet-50 model shows better results related with CNN pre-trained model. Deep CNN performing better for the end to end training process. For the COVID-19 detection more number of chest x-ray images can be evaluated in future and the various evolution stages can be analysed to help the radiologists in prediction [41]. This research also utilized the chest radiography images for an efficient COVID-19 prediction by the deep learning approaches. New CoroNet model developed in this study which is considered to be low cost and better results obtained. Higher sensitivity and accuracy resulted and thus this model is highly beneficial for the medical practitioners for proper understanding [42].

4.1. LSTM model

In public health system, strategic planning has been required to avoid deaths from COVID-19. The time series prediction of COVID-19 cases has been performed by LSTM, Bi-LSTM, autoregressive integrated moving average- ARIMA and support vector regression- SVR in 10 major COVID affected countries. This study estimated by means of the root mean square error, r2-score indices and absolute error. In this study BI-LSTM outperforms the other algorithms and it obtains reduced RMSE and MAE values. For better planning and management Bi-LSTM has been considered as better pandemic prediction algorithm [42]. This work utilized the Canadian health authority and John Hopkins university public datasets for COVID-19 forecasting model based on deep learning models. For future COVID-19 cases forecasting this study used the Long short term memory- LSTM. The possible ending point of the COVID outbreak has predicted in this study as June 2020 and compared it with USA, Italy and Canada transmission rates [43]. Due to the rapid population growth, automatic disease detection considered as challenging one. However automatic disease detection can support doctors in diagnostics. LSTM is combined with CNN in this study and utilizing the X-ray images to automatically detect COVID-19. Better accuracy, sensitivity, specificity have been resulted from this proposed system. Rapid diagnosis by doctors has been made from this study [44].

The time series prediction in which the data are in iterative way obtained by LSTM model. More accurate outputs have been predicted and number of positive cases have reported by LSTM. Apart from Google trends data, other data sources can be combined like mass media, screening registers, social media information, environmental and climate factors. Global prediction is necessary in terms of time series assessment [45]. For variety of disease prediction, SEIR models have been applied and however overfitting occurs since lot of predictor variables have used. In this study several combination of techniques have been executed based on LSTM, XgBoost and K-means to forecasting the short term COVID-19 cases in USA. Among the past days and forecasting, similarity is evaluated in this study using the k means algorithm with XgBoost technique. K-means with LSTM shows larger accuracy as resulted [46].

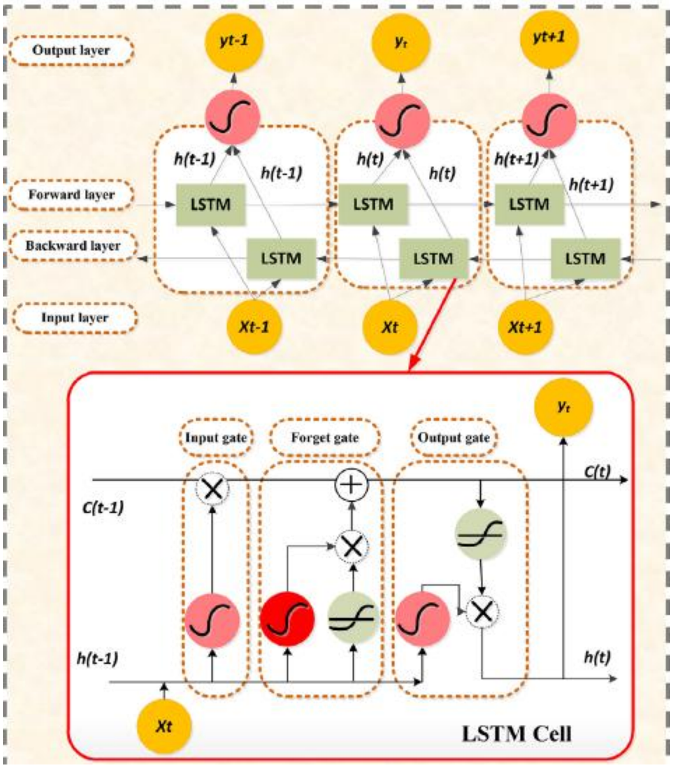

Fig.2. LSTM and Bi-LSTM architecture[42]

4.2. Bidirectional LSTM

With the intention of forecasting the cluster data based on COVID-19 Bi-LSTM model is established in this study. The prediction performance has improved which includes the lockdown information also [29]. The hospitalization estimation for coming week compared with present week has been inferred by the four recurrent neural network. Higher accuracy resulted in predicting the hospitalization in which every patient must receive suitable treatment. The hospitalization requirement ahs predicted before ad it has the potential to send warning message to the medical providers [47]. Various tweets have been found worldwide regarding with COVID-19 and these kind of tweets carries valuable information. It is highly challenging to process this information. To analyse the informative tweets, Bi-LSTM and other machine learning approaches are utilized for classification [48]. Various lockdown policies impact with respect to COVID-19 are evaluated and predicted in this study using the deep learning techniques. Various scenarios are evaluated related with lockdown policies and its effects are assessed while predicting COVID cases. The lifting of the lockdown especially for schools resulted in increases in infected cases simultaneously [49]. This research provided appropriate understanding of statistical growth rate of COVID cases in India. Most affected cases have been predicted using deep learning models [50].

5. CONCLUSION

COVID-19 is the major reason for infecting billions of people and affecting the economy worldwide. This study presented the detailed view of prediction and forecasting the COVID-19 cases worldwide. The forecasting models comprised with various deep learning models such as support vector regression (SVR), long shot term memory (LSTM), bidirectional long short term memory (Bi-LSTM) are assessed for time series prediction of confirmed cases, deaths and recoveries in ten major countries affected due to COVID-19. The paper also reviewed deep learning model to forecast the range of increase in COVID19 infected cases in future days. The comparative study also performed regarding the discussed deep learning models for COVID-19 prediction. This study provided the guidelines to the various other researchers who focusing the deep learning models in COVID-19 forecasting and prediction.

6. References

[1] B. Xu *et al.*, "Open access epidemiological data from the COVID-19 outbreak," *The Lancet Infectious Diseases,* 2020.

[2] W. Zhang, "Imaging changes of severe COVID-19 pneumonia in advanced stage," *Intensive care medicine,* pp. 1-3, 2020.

[3] S. Arik *et al.*, "Interpretable Sequence Learning for COVID-19 Forecasting," *Advances in Neural Information Processing Systems,* vol. 33, 2020.

[4] P. Nadella, A. Swaminathan, and S. Subramanian, "Forecasting efforts from prior epidemics and COVID-19 predictions," *European journal of epidemiology,* vol. 35, no. 8, pp. 727-729, 2020.

[5] F. Shan *et al.*, "Lung infection quantification of covid-19 in ct images with deep learning," *arXiv preprint arXiv:2003.04655,* 2020.

[6] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays," *Computer Methods and Programs in Biomedicine,* vol. 196, p. 105608, 2020.

[7] J. Zhang *et al.*, "Viral pneumonia screening on chest x-ray images using confidence-aware anomaly detection," *arXiv preprint arXiv:2003.12338,* 2020.

[8] L. Huang *et al.*, "Serial quantitative chest ct assessment of covid-19: Deep-learning approach," *Radiology: Cardiothoracic Imaging,* vol. 2, no. 2, p. e200075, 2020.

[9] K. Akrawinthawong, K. Majkut, S. Ferreira, and A. Mehdirad, "VOLTAGE-DEPENDENT INAPPROPRIATE RIGHT VENTRICULAR CAPTURE BY RIGHT ATRIAL LEAD PACING AS A CAUSE OF CARDIAC RESYNCHRONIZATION THERAPY NON-RESPONDER," *Journal of the American College of Cardiology,* vol. 69, no. 11S, pp. 2138-2138, 2017.

[10] M. J. Horry *et al.*, "COVID-19 detection through transfer learning using multimodal imaging data," *IEEE Access,* vol. 8, pp. 149808-149824, 2020.

[11] J. Civit-Masot, F. Luna-Perejón, M. Domínguez Morales, and A. Civit, "Deep learning system for COVID-19 diagnosis aid using X-ray pulmonary images," *Applied Sciences,* vol. 10, no. 13, p. 4640, 2020.

[12] M. H. D. M. Ribeiro, R. G. da Silva, V. C. Mariani, and L. dos Santos Coelho, "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil," *Chaos, Solitons & Fractals,* p. 109853, 2020.

[13] T. Chakraborty and I. Ghosh, "Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis," *Chaos, Solitons & Fractals,* p. 109850, 2020.

[14] M. Kamal, A. Aljohani, and E. Alanazi, "IoT meets COVID-19: Status, Challenges, and Opportunities," *arXiv preprint arXiv:2007.12268,* 2020.

[15] A. Rahman, M. S. Hossain, N. A. Alrajeh, and F. Alsolami, "Adversarial examples–security threats to COVID-19 deep learning systems in medical IoT devices," *IEEE Internet of Things Journal,* 2020.

[16] X. Ma *et al.*, "A survey on deep learning empowered IoT applications," *IEEE Access,* vol. 7, pp. 181721-181732, 2019.

[17] H. Jelodar, Y. Wang, R. Orji, and H. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach," *arXiv preprint arXiv:2004.11695,* 2020.

[18] A. S. Imran, S. M. Doudpota, Z. Kastrati, and R. Bhatra, "Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning--a Case Study on COVID-19," *arXiv preprint arXiv:2008.10031,* 2020.

[19] M. Haider Syed, S. Khan, M. Raza Rabbani, and Y. E. Thalassinos, "An artificial intelligence and NLP based Islamic FinTech model combining Zakat and Qardh-Al-Hasan for countering the adverse impact of COVID 19 on SMEs and individuals," 2020.

[20] Y. Guo *et al.*, "Innovating Computational Biology and Intelligent Medicine: ICIBM 2019 Special Issue," ed: Multidisciplinary Digital Publishing Institute, 2020.

[21] E. E.-D. Hemdan, M. A. Shouman, and M. E. Karar, "Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images," *arXiv preprint arXiv:2003.11055,* 2020.

[22] Q. Ni *et al.*, "A deep learning approach to characterize 2019 coronavirus disease (COVID-19) pneumonia in chest CT images," *European radiology,* vol. 30, no. 12, pp. 6517-6527, 2020.

[23] I. D. Apostolopoulos, S. I. Aznaouridis, and M. A. Tzani, "Extracting possibly representative COVID-19 Biomarkers from X-Ray images with Deep Learning approach and image data related to Pulmonary Diseases," *Journal of Medical and Biological Engineering,* p. 1, 2020.

[24] T. Kufel, "ARIMA-based forecasting of the dynamics of confirmed Covid-19 cases for selected European countries," *Equilibrium. Quarterly Journal of Economics and Economic Policy,* vol. 15, no. 2, pp. 181-204, 2020.

[25] M. Kolhar, F. Al-Turjman, A. Alameen, and M. M. Abualhaj, "A three layered decentralized IoT biometric architecture for city lockdown during COVID-19 outbreak," *IEEE Access,* vol. 8, pp. 163608-163617, 2020.

[26] J. Wang, H. Anh, F. Manion, M. Rouhizadeh, and Y. Zhang, "COVID-19 SignSym–A fast adaptation of general clinical NLP tools to identify and

normalize COVID-19 signs and symptoms to OMOP common data model," *ArXiv,* 2020.

[27] B. Tang, Z. Pan, K. Yin, and A. Khateeb, "Recent advances of deep learning in bioinformatics and computational biology," *Frontiers in genetics,* vol. 10, p. 214, 2019.

[28] H. Yi, S. T. Ng, A. Farwin, A. Pei Ting Low, C. M. Chang, and J. Lim, "Health equity considerations in COVID-19: geospatial network analysis of the COVID-19 outbreak in the migrant population in Singapore," *Journal of Travel Medicine,* 2020.

[29] A. B. Said, A. Erradi, H. Aly, and A. Mohamed, "Predicting COVID-19 cases using Bidirectional LSTM on multivariate time series," *arXiv preprint arXiv:2009.12325,* 2020.

[30] Y. Wang, M. Hu, Q. Li, X.-P. Zhang, G. Zhai, and N. Yao, "Abnormal respiratory patterns classifier may contribute to large-scale screening of people infected with COVID-19 in an accurate and unobtrusive manner," *arXiv preprint arXiv:2002.05534,* 2020.

[31] P. C. Silva, P. V. Batista, H. S. Lima, M. A. Alves, F. G. Guimarães, and R. C. Silva, "COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions," *Chaos, Solitons & Fractals,* vol. 139, p. 110088, 2020.

[32] N. Barak, U. Sommer, and N. Mualam, "Political Environment Aspects of COVID-19: Political Urban Attributes, Density and Compliance," *Density and Compliance (September 07, 2020),* 2020.

[33] W. J. Requia, E. K. Kondo, M. D. Adams, D. R. Gold, and C. J. Struchiner, "Risk of the Brazilian health care system over 5572 municipalities to exceed health care capacity due to the 2019 novel coronavirus (COVID-19)," *Science of the Total Environment,* p. 139144, 2020.

[34] A. Clark *et al.*, "Global, regional, and national estimates of the population at increased risk of severe COVID-19 due to underlying health conditions in 2020: a modelling study," *The Lancet Global Health,* vol. 8, no. 8, pp. e1003-e1017, 2020.

[35] M. Jonker, E. de Bekker-Grob, J. Veldwijk, L. Goossens, S. Bour, and M. Rutten-Van Mölken, "COVID-19 Contact Tracing Apps: Predicted Uptake in the Netherlands Based on a Discrete Choice Experiment," *JMIR mHealth and uHealth,* vol. 8, no. 10, p. e20741, 2020.

[36] A. Gupta, S. Banerjee, and S. Das, "Significance of geographical factors to the COVID-19 outbreak in India," *Modeling earth systems and environment,* vol. 6, no. 4, pp. 2645-2653, 2020.

[37] S. Chatterjee, A. Sarkar, M. Karmakar, S. Chatterjee, and R. Paul, "How the asymptomatic population is influencing the COVID-19 outbreak in India?," *arXiv preprint arXiv:2006.03034,* 2020.

[38] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *arXiv preprint arXiv:2001.05566,* 2020.

[39] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict covid-19 infection," *Chaos, Solitons & Fractals,* vol. 140, p. 110120, 2020.

[40] P. Arora, H. Kumar, and B. K. Panigrahi, "Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India," *Chaos, Solitons & Fractals,* vol. 139, p. 110017, 2020.

[41] A. M. Ismael and A. Şengür, "Deep learning approaches for COVID-19 detection based on chest X-ray images," *Expert Systems with Applications,* vol. 164, p. 114054, 2020.

[42] A. I. Khan, J. L. Shah, and M. M. Bhat, "Coronet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images," *Computer Methods and Programs in Biomedicine,* p. 105581, 2020.

[43] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos, Solitons & Fractals,* p. 109864, 2020.

[44] M. Z. Islam, M. M. Islam, and A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," *Informatics in Medicine Unlocked,* vol. 20, p. 100412, 2020.

[45] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R. N. Kalhori, "Predicting COVID-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study," *JMIR Public Health and Surveillance,* vol. 6, no. 2, p. e18828, 2020.

[46] S. R. Vadyala, S. N. Betgeri, E. A. Sherer, and A. Amritphale, "Prediction of the number of covid-19 confirmed cases based on k-means-lstm," *arXiv preprint arXiv:2006.14752,* 2020.

[47] Y. Meng, Y. Zhao, and Z. Li, "An early prediction of covid-19 associated hospitalization surge using deep learning approach," *arXiv preprint arXiv:2009.08093,* 2020.

[48] S. Chanda, E. Nandy, and S. Pal, "IRLab@ IITBHU at WNUT-2020 Task 2: Identification of informative COVID-19 English Tweets using BERT," in *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, 2020, pp. 399-403.

[49] A. B. Said, A. Erradi, H. Aly, and A. Mohamed, "A deep-learning model for evaluating and predicting the impact of lockdown policies on COVID-19 cases," *arXiv preprint arXiv:2009.05481,* 2020.

[50] A. Dutta, A. Gupta, and F. H. Khan, "COVID-19: Detailed Analytics & Predictive Modelling using Deep Learning," 2020.