STEM Fellowship Big Data Challenge 2021

CoAID: Detecting Misleading Information Using Deep Learning Models

Nghi Huynh[1]

[1]McGill University

May 31, 2021

Abstract

COVID-19 virus has rapidly spread around the world and affected lots of peoples' lives. Unfortunately, the diffusion of misinformation related to COVID-19 also gets created and propagates wildly on social media and other platforms. Since the spread of such misleading information has caused many adverse effects on public health, it is crucial to build computerized systems to understand, detect, and mitigate such misinformation. In this paper, we proposed a deep neural network for the detection of fake news. The deep learning models are the modified LSTM with one layer and the modified LSTM with two layers. In particular, we carried out our experiments with a large dataset from tweets and other platforms related to COVID-19. We then separated the dubious claims and news articles into two categories: fake and real. In addition, we set up our baselines with three machine learning models: logistic regression (LR), decision tree (DT), and support vector machines (SVMs). We then validated and compared the performances of these baselines with our LSTM models. The results obtained from our proposed models reveal high accuracy (89%) in distinguishing fake tweets from real tweets in the COVID-19 dataset. These results also show a significant improvement in our proposed model as compared to the existing state of art results from the baseline machine learning models. Our findings demonstrate the efficacy and accurateness of the LSTM models in detecting COVID-19 related misinformation and offer a methodological contributions to misinformation detection.

Keywords
infodemic, COVID-19, misinfomation detection, Deep Learning

1 Introduction

COVID-19 is believed to be caused by a coronavirus highly related to SARS-CoV-2. The emergence of this virus was first observed in Wuhan, China, in December 2019. Since then, COVID-19 has spread rapidly throughout the world. WHO declared the Public Health Emergency of International Concern outbreak on January 30, 2020 [1] and characterized COVID-19 as a pandemic on March 11, 2020 [2]. As of May 30, 2021, more than 170 million cases of COVID-19 and approximately 3.53 million deaths worldwide have been reported. Some common symptoms of COVID-19 include cough, trouble breathing, fever, sore throat, and loss of taste or smell [3].

While the COVID-19 scenario has gotten increasingly worse, social media with misleading information has risen and caused significant social disturbances. The spread of this misinformation has seriously affected our society. For example, fake cures such as Chloroquine led to the death of an Arizona man, or the 5G mobile network conspiracy destroyed 77 cell phone towers. Thus, the prevalent misinformation is not only threatening people's lives but also disrupting social order.

Misinformation is "the factually incorrect information that is not backed up with evidence" [4]. Misinformation is cast in different forms such as rumors, misleading content on the internet, and fake news. Besides, misinformation on social media has become an urgent and vital issue, especially in the health-related fields. There are many concerns related to COVID-19 misinformation since it can have adverse effects on public health.

Since then, infodemiologolists have developed many computational approaches to automatically and effectively detect COVID-19-related

misinformation [5, 6, 7]. However, this misleading information can spread exponentially in a day or two before being detected. When someone sees a piece of fake news multiple times, he/she tends to believe that it is true. Then, he/she starts to panic and spread the misleading information to more people. Therefore, those misleading information can go viral long before the intervention from the authority.

As harmful consequences from misinformation increase, early detection is crucial in a deterrence of such misinformation. In this paper, we built models and methods to identify those online misinformation. Since misinformation on social media are messages deliberately posted to persuade other users, we proposed a Deep Learning approach as opposed to a Machine Learning approach to detect those misleading messages. We generated some baseline machine learning models to compare the overall performances with our proposed deep learning models. Our Deep Learning model successfully recognized fake news with an 89% accuracy on the COVID-19 heAlthcare mIsnformation Dataset (CoAID). Our findings can be helpful to raise the public's awareness of the problems related to COVID-19 misinformation.

The rest of this paper is organized as follows. The Material and Methods section (Section 2) introduces the dataset and the methods dealing with feature extraction, data preprocessing, and the construction of our models. The Results section (Section 3) reports the experimental results for our proposed models and the Discussion (Section 4) section discusses the insights of these findings. Finally, the Conclusion section concludes our results and suggests further works.

2 Materials & Methods

2.1 Data Collection

We conducted our experiments using Twitter fake news dataset in CoAID (COVID-19 heAlthcare mIsnformation Dataset). CoAID has diversely confirmed fake and accurate news articles from either fact-checked or reliable websites and posts across social platforms [8]. The latest version of CoAID (Version 0.3) contains 278,385 news articles, 23,061 claims and ground truth labels. Since the dataset only contains tweetIDs and not the content of the tweet, we converted tweetIDs back into the content using a Hydrator tool. Then, we extracted only relevant features such as title or text to represent the content of the tweet and label features. The proposed flowchart of our study is shown in Figure 1.

2.2 Data preprocessing

Data preprocessing is an essential step in analyzing social media content for our detecting models. Since Twitter data is an unstructured dataset, we first refined it before applying any techniques. The preprocessing methods used in this work are lower-casing, punctuation removal, tokenization, stop-word removal, and lemmatizing. In our study, we applied these techniques as follows:

- Lower-casing: We converted each word in a text to lower cases such as "This is Blue" to "this is blue".
- Punctuation removal: We removed punctuation such as commas, apostrophes, quotes, question marks since they do not contain relevant information to our natural language model.
- Tokenization: We separated a piece of text into smaller units called tokens. The tokens can be further broken into words. For example, "the sky is blue" will be tokenized into ['the','sky','is','blue'].
- Stop-Word removal: We also removed stopwords from each tweet (i.e, articles, prepositions, conjunctions and some pronouns) since these words do not add much meaning to a sentence.
- Lemmatizing: Finally, we removed the suffix of a word and transformed it to its root word to reduce the number of word types of classes in our data. For example, the words "Playing", "Played", "Player" will be lemmatized to the root word "play".

2.3 Data preparation

We split the pre-processed data to 80% of the training set and 20% of the testing set. Then, we fed the training set into the ML/DL models, and we used the test set to validate the performance of our models.

2.4 Applying optimization and fitting learning models

We applied Machine Learning models as our baseline and Deep Learning model as our proposed approach. The details for each models are presented as follows:

2.4.1 Machine Learning models:

We applied three standard machine learning (ML) models after two steps. First, we used the

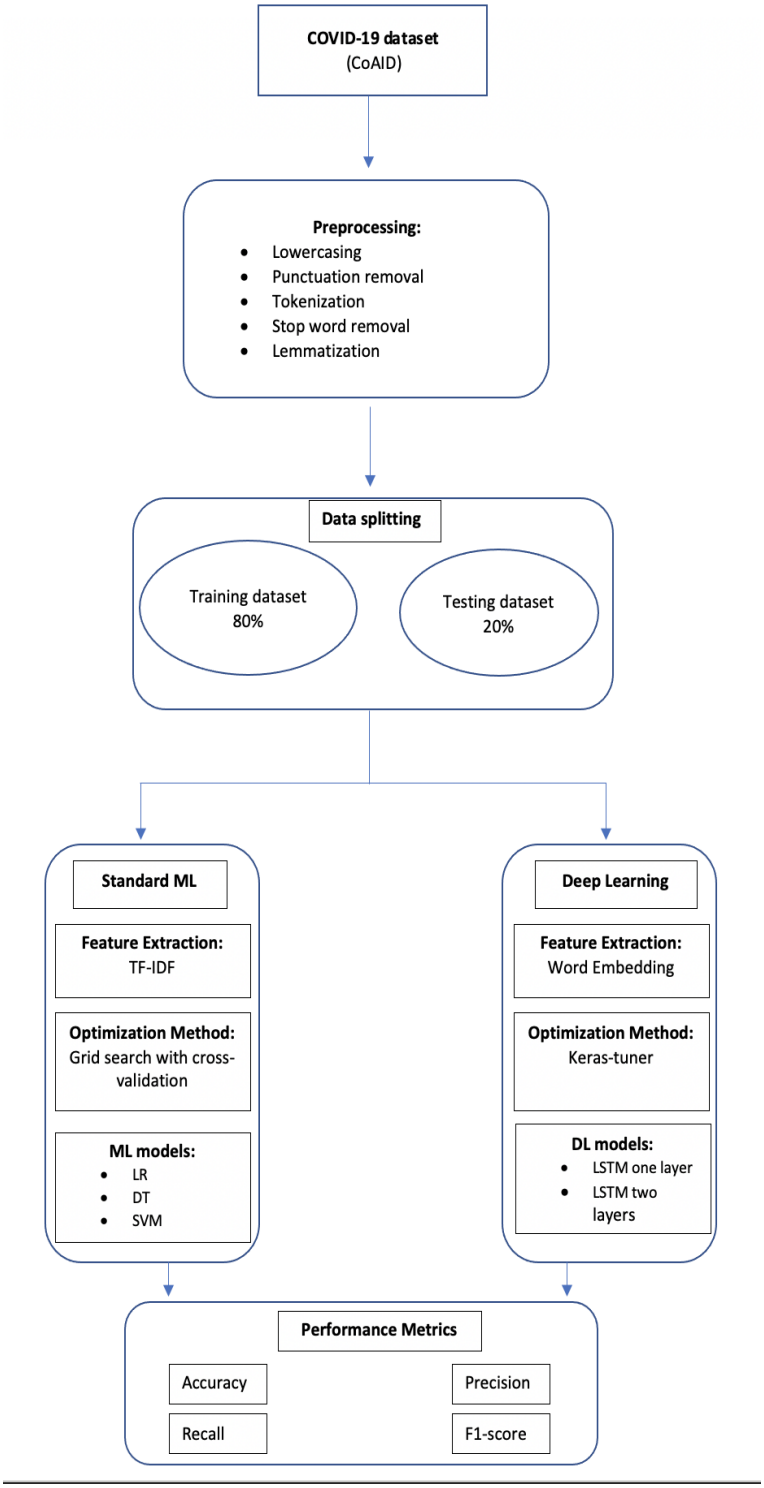

Figure 1: The proposed flowchart for detecting COVID-19 misinformation in CoAID

TF-IDF feature extraction method to represent sequences of words as vectors by assigning probabilities to sentences and lines of words. Then, we optimized the models using grid-search with cross-validation. The standard ML models are described as follows:

- Logistic Regression (LR): is a supervised ML model used in categorical classification [9]. This model is based on the sigmoid function in which the output ranges from 0 to 1. This model classifies the output as follows: the output is in group 0 if it is less than 0.5; otherwise, it is in group 1.
- Decision Tree (DT): is a non-parametric su-

pervised ML model used for classification and regression [10]. This model classifies the value of a target variable by learning simple decision rules inferred from the data features.

- Support Vector Machines (SVMs): are a set of supervised ML models for classification, regression, and outlier detection [source]. These models distinctly classify the data points by finding a hyperplane in an N-dimensional space (N-the number of features) [11].

2.4.2 Deep Learning model:

We applied our deep learning (DL) model after two steps. First, we used the word embedding feature extraction from the Word2Vector model to represent each word from a text numerically. Then, we optimized our model by tuning the number of neurons and dropout rate. Our DL model is described as follows:

- Long Short-Term Memory (LSTM): is a deep recurrent neural network used in classifying, processing, and making predictions based on sequence data. This network is more reliable than the tradidional recurrent network since it has feedback connections and allows longer time lags [12].

2.4.3 Evaluation models:

We determined the performance of our model in comparison to the baseline models based on the following statistics:

- Accuracy: is a measure of correctly identified samples out of all the dataset.
  $Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100$
- Precision: is a measure of result relevancy.
  $Precision = \frac{TP}{TP+FP}$
- Recall: is a measure of how many correctly relevant results are returned.
  $Recall = \frac{TP}{TP+FN}$
- F1-score: is a harmonic mean of precision and recall.
  $F1 - Score = \frac{2*Precision*Recall}{Precision+Recall}$

3 Results

This section described the overall performance of the three baseline ML models (LR, DT, SVMs) and our proposed DL models (Modified LSTM with one layer and modified LSTM with two layers). We used TF-IDF feature extraction with a matrix size of 2339 for each machine learning model. On the other hand, we used word-embedding feature extraction with a matrix size of 2339 for each of our proposed models.

We trained the ML and DL models with 80% of the dataset and then tested with 20% testing data. The three standard ML classifiers were implemented using the sci-kit-learn package in Python 3. The DL models were implemented using TensorFlow and Keras package in Python 3. For the hyperparameters of our DL models, we tuned the number of neurons and the dropout rate. The best values of these hyperparameters are shown in Table 1 .

DL Models	Neurons	Dropout
LSTM one layer	150	0.25
LSTM two layers	200	0.60

Table 1: The best hyperparameter values for the modified LSTM models

We then evaluated the performance of three baseline ML models and our DL models over the dataset. The results of cross-validation are shown in Figure 2. For the baseline models, the LR model obtained the highest efficiency with 85.4% of accuracy, 84.8% of precision, 86% of recall, and 85.4% of F1-score. Similarly, the SVM model got 84.6% of accuracy, 83.7% of precision, 83% of recall, and 84.95% of F1-score. The DT model obtained the lowest efficiency with 83% of accuracy, 87% of precision, 83% of recall, and 84.95% of F1-score. For our proposed models, the LSTM with two layers obtained the highest efficiency with 93.1% of accuracy, 94.59% of precision, 93.1% of recall, and 93.84% of F1-score. On the other hand, the LSTM with one layer got 92.84% of accuracy, 92.8% of precision, 92.4% of recall, and 92.6% of F1-score.

The results of testing are shown in Figure 3. For the baseline ML models, the testing results show that the LR model yielded the highest efficiency (accuracy of 82.4%, precision of 82%, recall of 85%, and F1-score of 84%), whereas the DT model obtained the lowest efficiency (accuracy of 80%, precision of 86%, recall of 78%, and F1-score of 83%). The SVM got 82% of accuracy, 82% of precision, 84% of recall, and 83% of F1-score. Regarding our deep learning models, the modified LSTM with two layers achieved the highest efficiency (89.32% of accuracy, 89.33% of precision, recall, and F1-score), whereas the modified LSTM with one layer only obtained 88.03% of accuracy, 88.08% of precision, recall, and F1-score.

Cross-validation results for the CoAID							
Feature Extraction Method	Models		Matrix Size	Measure Performance Methods			
				Accuracy	Precision	Recall	F1-Score
TF-IDF	ML	LR	2339	85.4±0.1	84.8±0.15	86±0.1	85.4±0.11
		DT		83±0.2	87±0.3	83±0.2	84.95±0.15
		SVM		84.6±0.09	83.7±0.15	84.2±0.08	83.95±0.16
Word Embedding	DL	LSTM one layer		92.84±0.14	92.8±0.15	92.4±0.14	92.6±0.07
		LSTM two layers		93.1±0.16	94.59±0.17	93.11±0.2	93.84±0.08

Figure 2: The performance of ML and DL models for the cross-validation results

Testing Results for the CoAID							
Feature Extraction Method	Models		Matrix Size	Measure Performance Methods			
				Accuracy	Precision	Recall	F1-Score
TF-IDF	ML	LR	2339	82.4	82	85	84
		DT		80	86	78	83
		SVM		82	82	84	83
Word Embedding	DL	LSTM one layer		88.03	88.08	88.08	88.08
		LSTM two layers		89.32	89.33	89.33	89.33

Figure 3: The performance of ML and DL models for the testing results

4 Discussion

From the results obtained above, we visualized it with the big picture for the cross-validation performances and the testing results in Figure 4 and Figure 5. We observe that the modified LSTM models have obtained the best performances in both cross-validation and testing results compared to the baseline ML models. Specifically, the modified LSTM with two layers has achieved the best accuracy (93.1%) for the cross-validation results. Moreover, the LSTM with two layers has yielded the highest precision (94.59%) and recall (93.11%). On average, the modified LSTM models have obtained the best testing performance in comparison to the baseline machine learning models. For the testing results, the modified LSTM with two layers has the highest accuracy of 89.32%, precision, recall, and F1-score of 89.33%. These results suggest that our LSTM models are the most effective and accurate ones in detecting fake news from the dataset.

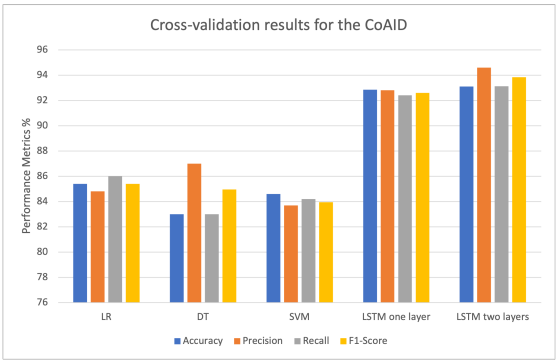

Figure 4: The cross-validation performances results for the CoAID

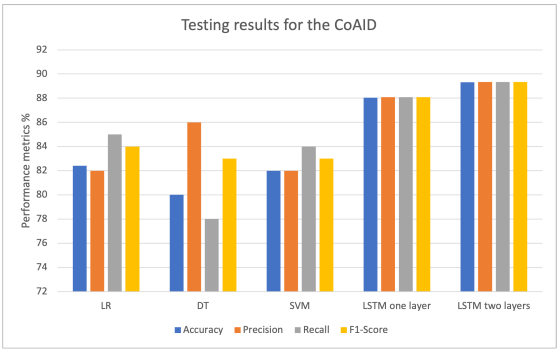

Figure 5: The testing results for the CoAID

On the other hand, we observe that the LR model has yielded the best performance among the standard ML models in cross-validation and testing results. Even though the DT model has got the lowest performances (accuracy of 83%, precision of 87%, recall of 83%, and F1-score of 84.95%), this model has achieved the highest precision (87%) among other baseline models. These results suggest that LR is the best and simplest model that can be applied to detect misinformation compared to other baseline models.

Consequently, our modified LSTM models for the CoAID outperformed other baseline models (LR, SVM, and DT) in both cross-validation and testing results. Based on these results, we suggest that the LSTM models are suitable for detecting fake news from CoAID and can be further applied in detecting misinformation from other datasets.

Conclusions

Due to the spread of COVID-19 misinformation on social media and other platforms, we presented the optimized baseline machine learning models and proposed efficient and enhanced deep learning models to detect misinformation for COVID-19 based on CoAID. We used cross-validation and testing to support the validity of our models. Regarding the baseline ML models, the LR obtains the highest performance in the testing results (accuracy: 82.4%, precision: 82%, recall: 85%, and F1-score: 84%). As for our proposed DL models, the LSTM (two layers) obtains the best testing results with the following performance metrics: the accuracy is 89.32%, the precision, recall, and F1-score are 89.33%. Thus, we can conclude that our proposed modified LSTM with tuned hyperparameters outperformed other baseline ML models (LR, DT, and SVMs).

Despite the contributions of this study, it also has several limitations. First, this study adopted the dataset (CoAID) generated from news articles, user engagement, and social platform posts, the label of the misinformation heavily relied on the performance of the dataset's algorithms. Thus, misclassification can be possibly carried on in our analysis. In future studies, we would consider developing a cross-checking annotation system before analyzing any dataset. Second, in this study, we explored and analyzed the features of misinformation based on previous literature [7]. Future works could explore and extend our research with other features to capture its dynamic in distinguishing misleading information from legitimate information.

Acknowledgements

We would like to thank the STEM Fellowship team for hosting this event. We would also like to extend our gratitude to the partners and sponsors for their contributions to the Undergraduate Big Data Challenge 2021.

References

[1] World Health Organization. Statement on the second meeting of the international health regulations (2005) emergency committee regarding the outbreak of novel coronavirus (2019-ncov).

[2] World Health Organization. Who director-general's opening remarks at the media briefing on covid-19-22 march 2020.

[3] Centers for Disease Control and Prevention. Symptoms of coronavirus. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html.

[4] Bode L and Vraga E. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4):619–638, 2015.

[5] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xu, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi modal fake news detection. *In Proceedings of the 24th acm sigkdd international conference on knowledge discovery and data mining*, pages 849–857, 2018.

[6] Natali Ruchansky, Sungyong Seo, and Yan Liu. A hybrid deep model for fake news detection. *In Proccedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.

[7] Diaa Salama Abdelminaam, Fatma Helmy Ismail, Mohamed Taha, Ahmed Taha, Essam H Houssein, and Ayman Nabil. CoAID-Deep: An optimized intelligent framework for automated detecting COVID-19 misleading information on Twitter. *IEEAcces*, 9:27840–27867, 2021.

[8] Limeng Cui and Dongwon Lee. CoAID: COVID-19 healthcare misinformation dataset, 2020.

[9] F.E. Harrell. "Ordinal logistic regression" in Regression Modeling Strategies. *Springer*, pages 311–325, 2015.

[10] J. R. Quinlan. Induction of decision tress. *Mach. Learn*, 1(1):81–106, 1986.

[11] C-W. Hsu, C.-C. Chang, and C.-J Lin. A Practical Guide to Support Vector Classification, 2003.

[12] F. A. Gers, J. Schmidhuber, and F. Cummins. *Learning to Forget: Continual Prediction With LSTM.* Edison, NJ, USA: IET, 1999.