Detecting Accredited & Non-accredited Online Pharmacies with Deep Learning

Haley Egan (vkb6bn), Seth Galluzzi (vzw6yk), Tulsi Ratnam (tr9sq), Mani Shanmugavel (fdf7gn)

Abstract

The objective of this study was to harness neural networks to solve the rampant problem of non-accredited online pharmacies performing unapproved sales of drugs. A BERT model and MobileNetV2 model were used to classify the text and images of online pharmacies as accredited and non-accredited. Both models achieved moderately high accuracy on the validation datasets, with the BERT model at 90% and the MobileNetV2 model at 87% accuracy. Future work using a MultiModal model is recommended to combine the image and text models for improved overall accuracy.

Motivation

The National Association of Boards of Pharmacy claims that 96% of all online pharmacies are illegitimate. Currently, the process of verifying legitimacy is manual and slow. In order to flag suspicious online pharmacies that may be selling unsafe medications, it is necessary to create an automated process that can detect whether an online pharmacy is accredited or not. There has yet to be a Deep Learning study that identifies non-accredited pharmaceutical websites. Our study is the first to train complex neural networks to detect accredited and non-accredited pharmaceutical websites.

Method

We implemented two types of models during our analysis. The first model analyzes the text of each website. The second model analyzes the images of each website. Our custom dataset includes web-scraped content from the home pages of 72 known accredited and non-accredited pharmacies. Cleaning the data and preparing it for text and image preprocessing was necessary before building the models. The scraped websites' html were parsed and transformed into readable text by eliminating stopwords, removing characters, and creating uniform spacing and case-size. This final corpus was used to create word embeddings, which were then vectorized and encoded. The encoder reads the entire sequence at once, maintaining context and meaning, so that the decoder can make accurate predictions from our model.

After cleaning and preprocessing the text data, we implemented a BERT architecture to analyze the text of each website. A BERT model generates a contextual representation of words. The existing pre-trained models were essential in helping us customize its architecture for our own dataset. BERT is renown for its performance in task-specific binary text classification problems.

We then built an image classification model from each pharmacy website's extracted images. The FDA requires legal online pharmacies to display valid certifications and licenses. In our dataset, the accredited websites have certifications and licenses, whereas

many of the non-accredited sites do not. The goal of the image model is to learn the images and to distinguish between accredited and non-accredited websites based on these image differences. The model may also find other differences between the images from the two classes of websites that are less apparent to the human eye.

A pipeline was created to scrape images from each website, and then sort them into training, validation, and test sets with 2 classes, accredited and non-accredited. The images were preprocessed by resizing them into the correct dimensions, and then normalized in order to feed into the model. In order to determine which image model performed best on our data, a variety of image classification models were tested, including VGG, CNN, ResNet, Xception, and MobileNetV2. Results indicated that the MobileNetV2 model had the best performance. MobileNet uses a convolutional neural network (CNN) architecture and is pre-trained for image classification problems.

Experiments

BERT Model

A BERT model was built for learning the text scraped from pharmacy websites. After the header, footer, and body text was scraped from the known accredited and non-accredited websites, the text for each website was inserted into a dataframe. A label column was created to identify each row (website) as either accredited (0) or non-accredited (1). The text was then combined into a single string for each row, split into training and validation sets, and cleaned using text preprocessing techniques. The Keras Universal Sentence Encoder, which can process 100 different languages, was used for the preprocessor and encoder. After the preprocessing stage, the BERT model was defined, compiled, and run. When building the BERT model, epochs, optimizer, loss, and dropout layer and activation layers were all tested and fine-tuned to achieve the best results.

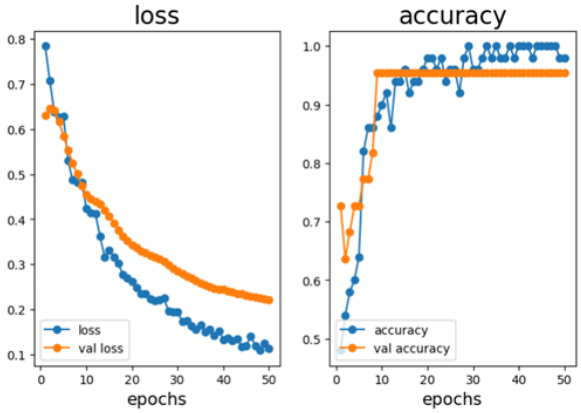
Several tests were run to find the optimal number of epochs for best results from the BERT model. An epoch is an iteration over the entire data provided. The model stops training when the number of epochs specified is reached. After testing with 20 epochs, it was apparent that accuracy was still improving each epoch, and would continue improving with an increased number of epochs. The number of epochs was then increased to 100. After about 45 epochs, the level of accuracy no longer improved at a significant rate. Therefore, it was determined that 50 epochs was the ideal number of epochs for the BERT model to reach the most accurate results. The optimizer helps ensure that the appropriate weights and learning rate are used in order to reduce losses, and helps increase the speed of model training. Several optimizers were tested to determine which performed best with the BERT model. 'Nadam', 'SGD', 'RMSprop','Adagrad' and 'Adam' were all tested. The 'Adam' optimizer was chosen as the best optimizer. 'Adam' is a stochastic gradient descent method based on the adaptive estimate of first-order and second-order moments.

The loss function is used to compute the quantity that a model should try to minimize during training. Since the BERT model is classifying pharmacy websites based on two label classes, either accredited (0) or non-accredited (1), the Keras loss function 'categorical_crossentropy' was used. In order to use this loss function, the labels are represented with one_hot encoding. For optimal results, the BERT model contains a dropout layer to help prevent overfitting during model training. The dropout rate was set at 0.2, which is the fraction of the input units to drop. Dropout rates of 0, 0.2, and 0.5 were tested, with a rate of 0.2 resulting in the best performance.

A dense layer was implemented for the BERT model. In a dense layer, each neuron in the layer receives input from all neurons in its previous layer. In this step,

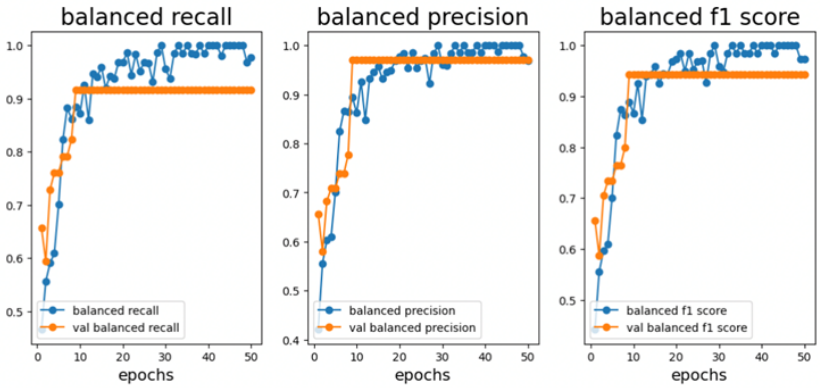
matrix-vector multiplication is performed, where the parameters are trained and updated with backpropagation, with a vector as output. Softmax activation was used for the last layer in the BERT model, which transforms the vector output from the dense layer into a vector of probabilities, for a probability distribution. In order to determine the success of the BERT model in classifying accredited and non-accredited online pharmacies, several metrics were evaluated, including: loss, accuracy, true negatives, false positives, AUC, recall, precision, and F1 score.

Figure 1. BERT Loss and Accuracy



The BERT model was able to correctly classify the validation data with 95% accuracy, and close to 100% accuracy for the training data. The training data loss was 12.56% while the validation data loss was much higher at 27.21% .

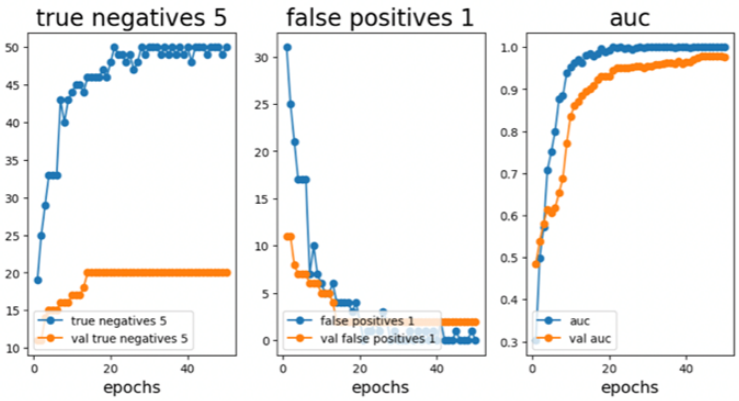
Figure 2. BERT Metrics



The results for recall, precision and F1 score were all very similar. For the training data, all three were close to 100%.For the validation data, the recall achieved 91% the f1-score was 94% and the precision score was 97% .

For this dataset, it is important to examine the false negative and false positive rates. It could be very harmful for the model to wrongly classify a non-accredited online pharmacy as accredited. This could affect people's wellbeing, who could trust a pharmacy that may be selling fake or expired medicines, or committing other types of scams. Therefore, it is important to add a greater weight to false positives. The training data performed better than the validation data when detecting true negatives and false positives. However, both performed fairly well overall.

Figure 3. BERT Metrics 2



The AUC-ROC curve is a good performance metric for the classification of pharmacies by the BERT model. ROC is a probability curve and AUC represents the measure of separability. The closer the curve is to 1, the better the BERT model is at distinguishing between accredited and non-accredited online pharmacies. The AUC-ROC curve for training data reached 1 after about 20 epochs, and .9773 after 40 epochs for the validation data. These are very promising results of the overall effectiveness of the BERT model.

After completing the basic BERT model that classified pharmacies by label (accredited and non-accredited), we had planned to build a multi-label BERT classification model. FDA regulations state that an accredited online pharmacy must provide a valid phone number and address on their website. The goal was to add the scraped address and phone numbers, run them through verification APIs, and label each website as having a valid address (0) or not (1), and having a valid phone number (0) or not (1). The multi-label BERT model would be able to learn from the multiple labels and different text inputs.

Unfortunately, we could not complete this part of the project due to several logistical challenges. For example, while most accredited websites did have the necessary information, it was sometimes not on the home page, which was the page we web-scaped. In order to obtain this information, web crawling through every page of the website would be necessary. Additionally, while the verification APIs worked, it was still difficult to prove the validity of the information. For example, a non-accredited website might have a valid phone number, but it might not belong to the pharmacy. It would require calling the number to confirm that it went through to an actual pharmacy. This next step of the project would be valuable to explore in future work, and would likely improve the overall success and accuracy of the classification process.

Image Model

In order to choose the best image model for our data, five image models were tested for highest accuracy. The classes (accredited or non-accredited) for the images were not one-hot encoded, thus our training and validation sets were all one-dimensional arrays. This helped us determine the loss function needed for each of the image models. We implemented the Sparse Categorical Cross-Entropy loss function with an Adam optimizer.

Several different image models were tested with the pharmacy dataset in order to determine which would work best for this study. For the VGG model, we implemented early stopping to prevent overfitting the model to the training data. VGG was trained with the lowest epoch of 10, and the CNN model had the highest number of epochs at

300. ResNet, Xception, and MobileNet all used 50 epochs. These last three models had the most gradual decline in both training and validation losses. However, the validation accuracy for ResNet was much more erratic than the other two.

ResNet produced the lowest accuracy on the validation set, at 79.27% and VGG was in the middle at 84.15% . CNN and Xception models had the same accuracy of 86.59% though the CNN was trained for a longer amount of time. MobileNetV2 performed the best, having the highest accuracy on the validation set (87.8%), a lower training time, and a more gradual decrease in loss. Typically, ResNet is known to perform better because it uses a deeper and more complex network to achieve high accuracy. MobileNet, on the other hand, is a smaller model and has a less complex network but retains accuracy by reducing the number of parameters or dimensions in the convolution layers.

Figure 4. Image Model Accuracy Scores

	Model	Accuracy
0	VGG	0.841463
1	CNN	0.865854
2	ResNet	0.792683
3	Xception	0.865854
4	MobileNetV2	0.878049

MobileNetV2

Once we selected MobileNetV2 as the best image classification model, we explored using different optimizers to improve the accuracy metrics. We re-trained the baseline MobileNet + Adam model with 25 epochs, resulting in a 89.02% accuracy on the validation set. This produced better results than when running the preliminary model with 50 epochs. In the below graph, the loss function for the validation set starts to increase again after about 30 epochs, indicating that the model did not represent the validation set as well as the training set. After reducing the number of epochs to 25, the variance between the training and validation loss decreased, resulting in a more optimal accuracy metric. The table on the right highlights the other optimizers used; each model ran for 25 epochs.

	Optimizer	Accuracy
0	Adam	0.890244
1	AdaGrad	0.878049
3	NAdam	0.878049
4	RMSProp	0.878049
2	Adamax	0.865854
6	Nestrov	0.853659
5	Momentum	0.841463

Figure 5. Optimizer Accuracy Scores

Figure 6. MobileNetV2 Accuracy & Loss

It was evident after running these models that the initial MobileNet + Adam model had the best performance. We created a confusion matrix for this model, which showed 39 true positives and 34 true negatives. This means the model was able to correctly classify 39 websites as accredited and 34 sites as non-accredited with 89% accuracy based on the images alone. Other metrics were also calculated, shown in the table below. The precision for classifying a website as accredited (0) was higher, at 97% than classifying as non-accredited (1), at 81% . Whereas the recall for identifying non-accredited (1) sites was higher, at 97% than accredited (0), at 83% . The overall f1-score was slightly higher for classifying accredited sites than non-accredited, indicating that the model is better at identifying accredited websites.

Figure 7. MobileNetV2 Metrics

	precision	recall	f1-score	support
0	0.97	0.83	0.90	47
1	0.81	0.97	0.88	35
accuracy			0.89	82
macro avg	0.89	0.90	0.89	82
weighted avg	0.90	0.89	0.89	82

Results

For this study, we built two individual models that learned the different aspects of a websites' content, the text and images. While a direct comparison of the two models cannot be made since they each perform tasks using different inputs, the results from both of the models indicate that a text-based model is better at identifying accredited and non-accredited pharmacy sites.

One of the advantages of the BERT model is that it is a bi-directional, context-based model. This means that the model is able to distinguish between words found in an accredited and non-accredited site based on the contextual meaning of the words. BERT has been used for scam detection and other binary classification problems, so we had expected the model to perform well.

The BERT model has an overall accuracy of 95% with an overall precision metric of 97% indicating that its ability to correctly classify a site as accredited is extremely high and the number of false positives is low. The recall metric is slightly lower, at 91% meaning that the number of false negatives is slightly higher. A false positive is when the model misclassifies non-accredited sites as accredited, whereas a false negative is

when the model misclassifies accredited sites as non-accredited. In this case, having a low false positive and a high precision metric is conducive to our goal in flagging non-accredited sites.

The MobileNetV2 model performed better than expected. MobileNet is a more condensed version of a deep convolutional neural network. Its architecture is unique in that it uses depthwise separable convolutions that reduces parameters and computation speed, while retaining high accuracy.

While the image model did not perform as accurately as the BERT model, it still gave an overall accuracy of 89% . This is fairly impressive, given that the model does not have any context other than learning from the images. The recall for BERT is only slightly higher than MobileNet, meaning the ratio of true positives to false negatives is fairly similar. MobileNet has a precision and recall metric of 90% and 89% respectively. This means the false positive rate is only slightly higher than the false negative rate. Ideally, we would want a much lower false positive rate since misclassifying a non-accredited site as accredited could have far-reaching consequences.

Figure 8. Model Comparison Metrics

Model	BERT	MobileNetV2
Accuracy (%)	95	89
Precision (%)	97	90
Recall (%)	91	89
F1 (%)	94	89

A future goal to continue this study would be to combine these models into a multi-modal model that would classify accreditation based on text and images, giving the model a more holistic representation of a website than just one feature.

Conclusion

The purpose of this study was to build a model that would detect whether an online pharmacy was accredited or non-accredited based on website content. We built two types of models that separately analyzed the text and images on the homepage of each website. The BERT model analyzed the text from each website and was able to correctly classify accredited and non-accredited pharmacies with 90% accuracy on the validation dataset. The MobileNetV2 model analyzed the images from each website and was able to correctly classify accredited and non-accredited pharmacies with 89% accuracy.

Our initial aim was to combine these two models into a multi-modal model that would concatenate the outputs from the text and image models and assess these features together to determine pharmacy legitimacy. However, we encountered several issues during this final step. Since we built two different models, we had separated the data by text and images. Our dataset included rows for each website's text, and each website had several image urls formatted as a list of lists. When building our image classification models, we subset the images from the main dataset such that each image was in its own row to download and separate them for training and validation sets. This resulted in a text dataset with 72 rows, an accredited website image dataset with 495 rows, and a non-accredited website image dataset with 738 rows.

When we attempted to implement the Keras multi-modal model, so that the outputs from each individual model would be an input in the multi-modal model, we found that the outputs had different dimensions and concatenating them after running them

independently did not work. The output from the image model was much larger than the text model.

Upon researching several solutions, the best method for combining these models was to take the output layer from the image model as a vector representation for each website and merge it with our text dataset. This way, each website has a unique text and aggregated image vector that can be fed into a multi-modal model. While we weren't able to develop a multi-modal model during this study, the results we derived from our separate models could be used to build one in the future.

The implications of our study are a step in the right direction to identifying and removing non-accredited pharmacies. Companies like Google are working with the FDA to remove pharmaceutical websites that the FDA flags as illegitimate. However, the FDA is ill-equipped to tackle the magnitude of this issue, because the number of new non-accredited online pharmacies outweigh the manual flagging of websites by the FDA. In order to flag suspicious online pharmacies that may be selling unsafe medications, it is necessary to implement an automated process that can detect whether an online pharmacy is accredited or not.

A solution like ours could become a tool for government organizations like the FDA, or public companies like Google, to quickly identify and remove non-accredited online pharmacies. By removing the manual process of determining whether or not an online pharmacy meets legal requirements, the automatic detection of these websites could keep up with the fast-paced and ever-changing internet. This is the only way to prevent the scams and potentially harmful medical implication of non-accredited online pharmacies. Our study may also be used as an example on how to approach other internet scam detection issues.

Member Contributions

Tulsi - worked on web scraping and looping through the accredited websites to put into a dataframe, created a pipeline to download and sort images for image modeling, wrote parts of the project milestone, wrote final report, presented project proposal, presented final presentation
Seth - scraped the initial website lists from the NABP and FDA and worked on web scraping text and images from the accredited websites to put into a dataframe. Wrote pieces of project milestone 2, made final presentation
Haley - web scraped and parsed non-accredited websites, aggregated the accredited and non-accredited content into a complete dataset, built the BERT model, wrote literature review, project proposal, and final report. Gave presentation.
Mani - worked on scraping website urls from the non-accredited sites and converting it into a dataframe, filter zip codes and phone numbers from scraped data and validate the zip codes phone numbers using APIs, built the image classification models.

References

1. Github repository. https://github.com/HaleyEgan/Detecting-Accredited-vs-Unaccredited-Online-Pharmacies-with-Multimodal-Deep-Learning.

2. e. a. Ahmed Abbasi, Zhu Zhang. Detecting fake websites: The contribution of statistical learning theory. *MIS Quarterly*, 3, 2010. https://www.jstor.org/stable/25750686.

3. BeautifulSoup. Beautiful soup: Build a web scraper with python. *Real Python*, 2022. https://realpython.com/beautiful-soup-web-scraper-python.

4. J. Briggs. How to train a bert model from scratch. *Towards Data Science*, 2021. https://towardsdatascience.com/how-to-train-a-bert-model-from-scratch-72cfce554fc6.

5. Food and D. Administration. Internet pharmacy warning letters. *FDA*, 2022. https://www.fda.gov/drugs/drug-supply-chain-integrity/internet-pharmacy-warning-letters.

6. Fransiska. Differences between inception, resnet, and mobilenet. *Medium*, 2019. https://medium.com/@fransiska26/the-differences-between-inception-resnet-and-mobilenet-e97736a709b0.

7. P. Huilgol. Top 4 pre-trained models for image classification with python code. *Analytics Vidhya*, 2022. https://www.analyticsvidhya.com/blog/2020/08/top-4-pre-trained-models-for-image-classification-with-python-code/.

8. D. V. Kumar. Mobilenet vs resnet50 – two cnn transfer learning light frameworks. *Analytics India Mag*, 2020. https://analyticsindiamag.com/mobilenet-vs-resnet50-two-cnn-transfer-learning-light-frameworks/.

9. NABP. Accredited digital pharmacies. *NABP*, 2022. https://nabp.pharmacy/programs/accreditations-inspections/digital-pharmacy/accredited-digital-pharmacies/.

10. A. Nandan. Text extraction with bert. *Keras*, 2020. https://keras.io/examples/nlp/text_extraction_with_bert/#preprocess-the-data.

11. Swatimeena. Bert text classification using keras. *Medium*, 2020. https://swatimeena989.medium.com/bert-text-classification-using-keras-903671e0207d.

12. J. L. C. Zheng Dong, Kevin Kane. Detection of rogue certificates from trusted certificate authorities using deep neural networks. *ACM Transactions on Privacy and Security*, 2016. https://www.researchgate.net/publication/316940647_Detection_of_Rogue_Certificates_from_Trusted_Certificate_Authorities_Using_Deep_Neural_Networks.

[9] [5] [2] [12] [3] [10] [11] [4] [7] [6] [8] [1]