Model Design and Implementation

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

Abstract

In this study, we are using different forms of deep learning such as Natural Language Processing (NLP) and image recognition to build a multi-modal neural network to detect whether a pharmaceutical website is accredited or not accredited.

Motivation

According to the executive director of the National Association of Boards of Pharmacy, 96% of all online pharmacies are illegitimate [4]. 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 illegal pharmaceutical websites. Our study will be the first to train a complex neural network to detect accredited versus non-accredited pharmaceutical websites.

Literature Survey

Research conducted by the National Association of Boards of Pharmacy (NABP) and the U.S. Food and Drug Administration (FDA) was vital to our study. The NABP website has a list of accredited pharmacies and also outlines its methodology for verifying that a site is legitimate. Characteristics such as address, US telephone number, and state licenses displayed on the homepage are indicators that an online pharmacy is authentic. The non-accredited pharmacies were gathered from a list of internet pharmacy warning letters issued by the FDA. Warning letters are issued when online pharmacies violate the U.S. Federal Food, Drug, and Cosmetic Act. Violations include the offering of unapproved prescription drugs, offering of prescription drugs without a prescription, and offering of prescription drugs without FDA-required warnings. Furthermore, the research of Ahmed Abbasi, Zhu Zhang, David Zimbra, Hsinchun Chen and Jay F. Nunamaker Junior in Detecting Fake Websites: The Contribution of Statistical Learning Theory and the research of Zheng Dong, Kevin Kane, L. Jean Camp in Detection of Rogue Certificates from Trusted Certificate Authorities Using Deep Neural Networks also proved useful when exploring what data to collect from the list of pharmacies.

The initial model focuses on text classification - using the text on each websites' homepage to detect its legitimacy. Since our initial exploration of existing literature on text classification, many of which used NLTK and LSTM for models, we found that the BERT (Bidirectional Encoder Representations from Transformers) architecture was most relevant to our study. Lastly, we must recognize the Beautiful Soup Documentation, GeeksforGeeks, and stackoverflow for answering a multitude of questions along the way regarding the creation of our dataset. The trials, tribulations,

and errors throughout the data collection process would not have been overcome without the help of the data science community.

Method

Our custom dataset includes scraped content from the header, footer, and body of all accredited and non-accredited pharmacies. Cleaning the data and preparing it for preprocessing was necessary before we could start model building. 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.

One of the FDA requirements for accreditation is verified addresses and phone numbers. We used the APIs provided by USPS[1] and Verifone [2] to validate the zip codes and phone numbers for each website, and engineered the categorical variables 'zipcode' and 'valid_phone' in our dataset. The API "CityStateLookup" was used to validate the zip code. This requires an XML with the zip code as the request and outputs the city of the zip code as the response. Our code builds the XML request for the API using the python module "xml.etree.ElementTree". The Verifone API was used to validate the phone numbers. This API accepts the phone number as input and outputs a binary response of valid or not valid. These variables will be used as additional features in our final model.

After cleaning and preprocessing our data, we implemented a BERT architecture for our baseline model to analyze the text of each website. A BERT model generates a contextual representation of words and has been used for sentiment analysis and spam detection. 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. One of our goals is to predict whether a website is accredited or non-accredited just based on a websites' text and we believe this method is conducive to our goal. Will the model be able to use context clues from a website to see whether the content/services advertised are from a legitimate pharmacy?

Preliminary Experiments

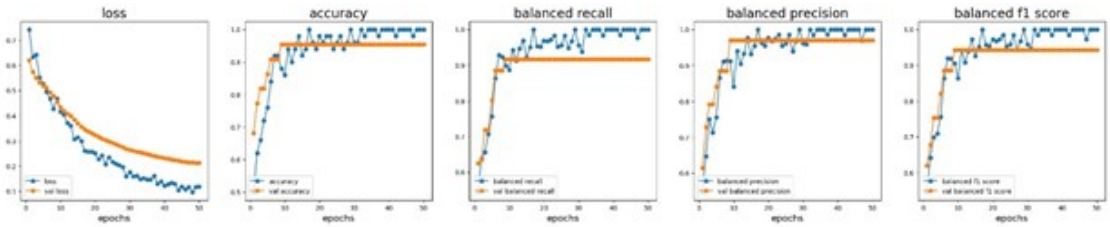
The current dataset contains a subset of 72 known accredited and non-accredited online pharmacies. Given the website text and classification label, accredited (0) and non-accredited (1), the model was able to correctly classify the test data with 95% accuracy, 97% precision, 91% recall, and a f1-score of 94%Ḟor optimal results, the BERT model contained a dropout and a dense layer, with softmax as the activation function. 50 epochs were run, 'Adam' was used as the optimizer, and 'categorical_crossentropy' was used as the loss.

After only 10 epochs, accuracy, recall, precision, and the f1 score all converged between 0.9 and 1.0. After about 30 epochs for the training data, all of the metrics plateaued close to 1.0. The testing/validation data consistently plateaued after 10 epochs. The loss continued to decrease for all 50 epochs.

[1] https://secure.shippingapis.com/ShippingAPITest.dll
[2] api.veriphone.io

Figure 1. Results from baseline model

```
Epoch 48/50
2/2 [==============================] - 7s 4s/step - loss: 0.0961 - accuracy: 1.0000 - balanced_recall: 1.0000 - balanced_precision: 1.0000 - balanced_f1_score: 1.0000 -
val_loss: 0.2136 - val_accuracy: 0.9545 - val_balanced_recall: 0.9167 - val_balanced_precision: 0.9706 - val_balanced_f1_score: 0.9429
Epoch 49/50
2/2 [==============================] - 7s 4s/step - loss: 0.1162 - accuracy: 1.0000 - balanced_recall: 1.0000 - balanced_precision: 1.0000 - balanced_f1_score: 1.0000 -
val_loss: 0.2130 - val_accuracy: 0.9545 - val_balanced_recall: 0.9167 - val_balanced_precision: 0.9706 - val_balanced_f1_score: 0.9429
Epoch 50/50
2/2 [==============================] - 7s 4s/step - loss: 0.1194 - accuracy: 1.0000 - balanced_recall: 1.0000 - balanced_precision: 1.0000 - balanced_f1_score: 1.0000 -
val_loss: 0.2123 - val_accuracy: 0.9545 - val_balanced_recall: 0.9167 - val_balanced_precision: 0.9706 - val_balanced_f1_score: 0.9429
```



Next Steps

The first model contained two columns, one feature ('text'), and one label (0/1), and contained 72 rows of websites. The results from this baseline will be used for comparison as we build onto our model. Our final dataset will contain 150 websites. Additionally, features and labels for addresses and phone numbers will be added to see if these will improve the model's accuracy, precision and recall metrics. The BERT model will become a multi-label classification model.

We are also in the process of building our image classification model. The scraped images from all of the websites will be sorted into training, validation, and test sets with 2 classes: accredited and non-accredited. We will be implementing a Convolutional Neural Network (CNN) to detect if our model correctly predicts whether a website is legitimate or not solely based on a websites' images. The accredited websites all have certifications and licenses whereas the non-accredited sites do not. Our hope is that the model will be able to pick up on this distinction to be able to correctly classify the websites. Finally, we will be combining both models using a Keras Functional API. This allows us to have multiple inputs which are then concatenated to produce a single output. The results from both the BERT and CNN models will be used as inputs in the Functional API.

Member Contributions

Haley: Wrote literature review and project proposal, parsed non-accredited websites, aggregated the accredited and non-accredited content into a complete dataset, built the BERT model.
Seth: Scraped the initial website lists from the NABP and FDA, worked on web scraping text and images from the accredited websites to put into a dataframe, wrote pieces of project milestone.
Tulsi: Presented project proposal, contributed to web scraping content, parsed accredited websites for usable urls, looped through accredited websites to convert into a dataframe, wrote parts of the project milestone.
Mani - Scraped website urls from the non-accredited sites and converted into a dataframe, wrote the code to validate the zipcodes and phone numbers using APIs from scraped data.

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

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

6. 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.

[2] [6] [3] [5] [4] [1]