# Detecting Accredited vs Non-accredited Online Pharmacies with Multimodal Deep Learning

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

## Introduction

As the use of the internet and social media have become ubiquitous, so too have the number of online scams, fraudulent behavior, and black market trends. One specific issue that has gone unsolved is the expanding number of non-authentic and non-accredited online pharmacies, many of which conduct illegal pharmaceutical sales and scams.

According to the executive director of the National Association of Boards of Pharmacy, 96% of all online pharmacies are illegitimate [6]. Identifying whether an online pharmacy is authentic or not is a major problem. Currently, the process of verifying legitimacy is manual and slow. Companies like Google are working with the Federal Drug Administration (FDA) to remove pharmaceutical websites that the FDA flags as illegal. However, some news sources have argued that the FDA is ill-equipped to tackle the magnitude of this issue, because the number of new unaccredited online pharmacies outweigh the manual flagging of websites by the FDA [12]. 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.

For this project, we propose a study that harnesses Multimodal Deep Learning, using different forms of deep learning such as Natural Language Processing (NLP) and image recognition to train models to detect whether a pharmaceutical website is accredited or not accredited. Many studies have been conducted that identify fake web articles and scam websites using Deep Learning. Some studies have been conducted to identify illicit drug sales online and through social media using NLP and image recognition. However, there have been no Deep Learning studies to identify illegal pharmaceutical websites.

Non-accredited online pharmacies is an ever growing issue on the internet that could have harmful consequences, such as the sale of fake or expired drugs, and identity theft or financial scams. Our study will be the first to train complex neural networks to detect accredited versus non-accredited pharmaceutical websites.

## Literature Review

### Detecting Scam vs Real Websites

There have been some Machine Learning studies conducted to detect scam websites. A 2010 study used Support Vector Machine (SVM) to classify whether a website is legitimate or not. This paper focuses on identifying fake websites that attack web users and perform identity theft. For the SVM model used in the study, the researchers collected features including the website text, source code, URL, and images [1]. This study is directly relevant to the study we are conducting. For our study we will collect similar features by scraping authentic and non-authentic websites. We will also have a mix of image and text data. However, while this study only used supervised learning

models, our study will use a combination of supervised and unsupervised learning with neural networks.

## Detecting Fake vs Real News Articles

Many more studies have been done using Deep Learning and Machine Learning to identify fake news articles on the internet. While this is not directly relevant to our research goals, as we will not be looking at news articles, these studies harness Natural Language Processing (NLP) to identify false claims. In our study, we will also perform NLP to detect non-authentic attributes and language.

One study, *Detecting Fake News with Natural Language Processing*, published on Analytics Vidhya, conducted NLP preprocessing with the NLTK library by removing stopwords, tokenization, and lemmatization to simplify and reduce the data. Then, vectorization was used to map words and phrases from vocabulary to numbers for predictions, and identifying similarities and semantics. Count Vectorizer, Hash Vectorizer, and TF-IDF Vectorizer were all used for preprocessing. These are all methods that can be used in our study. However, the Machine Learning models of Logistic regression, Naive-Bayes, Decision Tree, and Passive-Aggressive Classifier were used for predictions, with the Passive-Aggressive Classifier performing best [7]. In our study, we will train neural networks, rather than relying on supervised Machine Learning models.

*Fake News Everywhere: How to Detect It with SOTA NLP*, published on Towards Data Science, is another example of a study done to detect fake news articles. 'Fake' articles are labeled 1, and 'Real' articles labeled 0. XGB Classifier was chosen as the best model for predicting the correct labels [13]. While this may be a good resource for NLP preprocessing, it is not relevant for the Deep Learning algorithms we will be using. In the study, Predicting Fraudulent News Articles Using NLP + Deep Learning, text was transformed using Count Vectorizer and TF-IDF to obtain word counts and frequency scores. Model training was performed with a Multinomial Naive Bayes Classifier and a Logistic Regression Classifier to predict authenticity. LSTM neural network was used, with the maximum number of words used, and maximum number of words in each text. Results were analyzed using accuracy and AUC-ROC scores. Logistic Regression with Count Vectorizer performed the best [2].

## Detecting Illicit Online Drug Sale

The use of the internet and social media for illegal drug activities is not a new phenomena. One study was conducted in 2019 using Machine Learning to detect illicit drug dealers on Instagram. The researchers collected 12,857 Instagram posts using a web-scraper, and then compared their deep learning model against 3 supervised learning models, Random Forest, Decision Tree, and Support Vector Machine. The Deep Learning model used Long-Short-Term-Memory (LSTM) in the neural network to learn the pattern of texts associated with drug dealing posts. Their unsupervised Deep Learning model reached 95% accuracy on the F1 score, and performed better than the three supervised Machine Learning models [8].

A similar study was conducted to identify fraudulent COVID-19 products on social media, using NLP and Deep Learning. The study examined over 200,000 Twitter and Instagram posts during the initial wave of the COVID-19 pandemic to identify fraudulent products being sold, including test kits, medication, therapeutic methods, and a small number offering vaccines. A Biterm Topic Model (BTM) was used for tagging and labeling [9].

One study conducted to detect illicit drug ads on Google utilized SVM classification through TF-IDF preprocessing, as well as CNN classification through word embedding

preprocessing [14]. While the topic of these studies are not directly relevant to our study, the Deep Learning processes used could be applied to aspects of our study. TF-IDF and word embedding can both be applied to the preprocessing stage of our NLP modeling. SVM and CNN classifications can also be used for NLP in our study. However, BTM for tagging and labeling, as done in the COVID-19 study, will only be useful in our study if examining short strings. But there is likely a more applicable method, such as RNNs.

### Detecting Fake vs Real Certificates

For one aspect of our study, we will train our model to detect authentic vs non-authentic licenses on online pharmacy websites. One relevant study did something similar, using deep learning to detect real vs fake certificates. In this study, the researchers used Deep Neural Networks (DNN) and SVM to identify fraudulent certificates. The DNN model outperformed the SVM model [15].

## Experiments

### Data Collection

According to the National Association of Boards of Pharmacy, there are several red flags that can indicate if an online site is a scam: the site does not ask for a prescription, there is no way to speak to a pharmacist, no license, and fear-based tactics may be used [6]. The FDA says that an accredited pharmacy should always require a doctor's prescription, has a physical address and telephone number in the United States, is licensed in the state(s) in which they are operating, is licensed in all states in which they do business; and has a state-licensed pharmacist on staff to answer patient questions [5].

Since there is no existing dataset for this topic, we will build our own. The NABP has developed a website where users can check whether or not a website is legitimate, and provide a public list of all accredited online pharmacies [11]. We will use this list of 87 known accredited pharmacies to train our model to identify authentic pharmacies. The FDA provides a public list of 45 known non-authentic online pharmacies. For our study, we will use this list to train our model to identify non-authentic online pharmacies. Our ultimate goal is to pass a new URL into our system and identify if the website is accredited or not by using deep learning models to verify the requirements.

### Classification

From the FDA list of requirements necessary to be an accredited online pharmacy, we will focus on the following aspects: a physical address, a telephone number in the United States, a license in the state(s) in which they are operating or do business. We will also use NLP on the text within the websites to determine if there are any other unknown identifying factors.

After determining whether or not a website has the required items to be an accredited pharmacy, an additional step of authentication is required. Our models will not only need to determine if a website has the necessary requirements, but also that the items are authentic, requiring a two-step classification process. After determining through image recognition, whether an online pharmacy has a license (0) or not (1), our model will then need to determine whether the license is authentic (0) or not (1). The model will compare the extracted licenses against a library of known-authentic licenses that we provide. Similarly, after using NLP to determine whether the website has an address (0) or not (1), our model will verify with an API whether the address is authentic (0) or not (1). The same process will be the same for whether there is a

phone number (0) or not (1), and if it is authentic (0) or not (1). The final part will be to use NLP to determine if the model can use the body text of a website to determine if a pharmaceutical website is authentic (0) or not (1).

## Data Pipeline

## Extract, Transform, Load (ETL)

The first step in the pipeline will be to scrape the 87 known accredited pharmaceutical websites and the 45 known non-accredited pharmaceutical websites. The scraping will be done using the BeautifulSoup library [3]. All components of the websites will be scraped, parsed, and then added to the dataframe as individual features. Types of features will include: URL, website name/title, phone number, address, license images, and text.

One dataframe will be created for all of the data collected, with a feature distinguishing the known accredited websites from the known non-accredited websites. The next step will be to connect to APIs that can verify whether the scraped phone numbers and addresses are real or not. USPS provides APIs to validate a zip code or address, which will be used to validate the addresses of the pharmacies.[1] A number of APIs are also available online to validate phone numbers.[2] Free trials offered by these APIs will be leveraged for this project. Features will be added for each website distinguishing between real and fake addresses and phone numbers. For the text data, preprocessing will be performed to prepare the text for the training models. Once all scraping, parsing, cleaning, and loading is completed, modeling can be performed.

## Multimodal Deep Learning

For this study, we will use Multimodal Deep Learning (MDL) in order to harness multiple types of data. The main two types of data will be text and images. A neural network like a Convolutional Neural Network (CNN) would be appropriate for training on images. A neural network like Long-Short-Term-Memory (LSTM) would be appropriate for modeling text data. For MDL, these different neural networks can be combined through concatenation, and then applying softmax, in order to achieve a single output. Since the data are different and of different sizes, it is necessary to add weighted combinations to the subnetworks. This process can improve the overall outcome of the predictions when compared to modeling the neural networks separately [10].

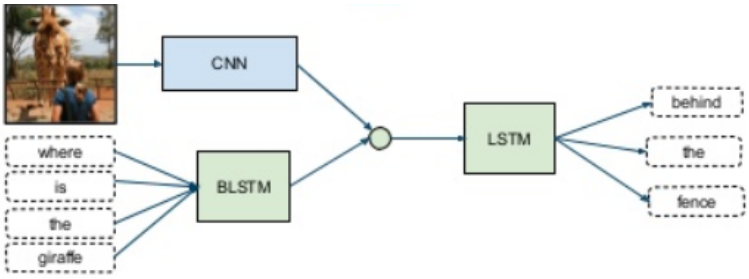


**Figure 1.** Example of Multimodal Model

[1] https://www.usps.com/business/web-tools-apis/address-information-api.htm
[2] https://veriphone.io/docs/v2 or https://www.neutrinoapi.com/api/phone-validate/

### Summary

Count Vectorizer, TF-IDF Vectorizer, and LSTM were commonly used for NLP, and accuracy, AUC-ROC scores, and F1 scores were used to analyze results in the studies we examined. The most relevant model we examined for image data are CNNs for unsupervised training. For our study, both of these will be combined in Multimodal Deep Learning to achieve optimal identification of accredited vs non-accredited online pharmacies.

## References


1. 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.

2. M. Asadullah. Predicting fraudulent news articles using nlp + deep learning. *Towards Data Science*, 2021. https://towardsdatascience.com/predicting-fraudulent-news-articles-using-nlp-deep-learning-ffdf64f19537.

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

4. P. G. Danny Valdez. Neutral or framed? a sentiment analysis of 2019 abortion laws - sexuality research and social policy. *SpringerLink*, 2022. https://link.springer.com/article/10.1007/s13178-022-00690-2.

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. M. Frellick. More illegal sites running online abortion pill scams. *WebMD*, 2022. https://www.webmd.com/women/news/20220804/illegal-sites-running-online-abortion-pill-scams.

7. K. Kajal. Detecting fake news with natural language processing. *Analytics Vidhya*, 2022. https://www.analyticsvidhya.com/blog/2021/07/detecting-fake-news-with-natural-language-processing/.

8. J. Li, Q. Xu, N. Shah, and T. K. Mackey. A machine learning approach for the detection and characterization of illicit drug dealers on instagram: Model evaluation study. *Journal of medical Internet research*, 2019. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6598421/.

9. T. K. Mackey, J. Li, V. Purushothaman, M. Nali, N. Shah, C. Bardier, M. Cai, and B. Liang. Big data, natural language processing, and deep learning to detect and characterize illicit covid-19 product sales: Infoveillance study on twitter and instagram. *JMIR public health and surveillance*, 6(3), 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7451110/.

10. P. Mehta. Multimodal deep learning: Fusion of multiple modalities using deep learning. *Towards Data Science*, 2018. https://towardsdatascience.com/multimodal-deep-learning-ce7d1d994f4.

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

12. R. Reader. The web is home to an illegal bazaar for abortion pills. *POLITICO*, 2022. https://www.politico.com/news/2022/08/01/the-web-is-home-to-an-illegal-bazaar-for-abortion-pills-the-fda-is-ill-equipped-to-stop-it-00048

13. M. A. Warsame. Fake news everywhere: How to detect it with sota nlp. *Towards Data Science*, 2021. https://towardsdatascience.com/fake-news-everywhere-how-to-detect-it-with-sota-nlp-f2dc1e07247c.

14. F. Zhao, P. Skums, A. Zelikovsky, E. Sevigny, M. Swahn, S. Strasser, and Y. Wu. Detecting illicit drug ads in google+ using machine learning. *Springer*, pages 171–179, 01 2019.

15. 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] [8] [14] [7] [2] [13] [5] [11] [12] [6] [1] [15] [3] [4] [10]