Detecting a Phishing URL by Machine Learning and NLP

---------*---------*---------*--------

Our Mentor: **Bipul Shahi Sir**

Team Members:

1. **Abhishek Mathur** (B. Tech IInd Yr. CSE) (Ideal Institute of Technology Ghaziabad)
2. **Hemant Yadav** (B. Tech IIIrd Yr. CSE) (Ideal Institute osf Technology Ghaziabad)
3. **Shubham Kumar Singh** (B. Tech Ist Yr. CSE) (Ideal Institute of Technology, Ghaziabad)

- **What is Phishing?**
  - **Phishing is an act of attempt to acquire information such as usernames, passwords, and credit card details, etc** of a person or organization illegally in an electronic communication.
  - Typically, **a victim receives a message that appears to have been sent by a known contact or organization**. The message contains malicious software targeting the user's computer or has links to direct victims to malicious websites in order to trick them into divulging personal and financial information, such as passwords, account IDs or credit card details.
  - In phishing the criminals creates a **fake website whose looks and feel are identical to the legitimate one**, in which the victims are told to enter their confidential details like username, password or account details.
  - Then Using this information criminal access that valid page like Facebook, your Bank website and do criminal tasks.

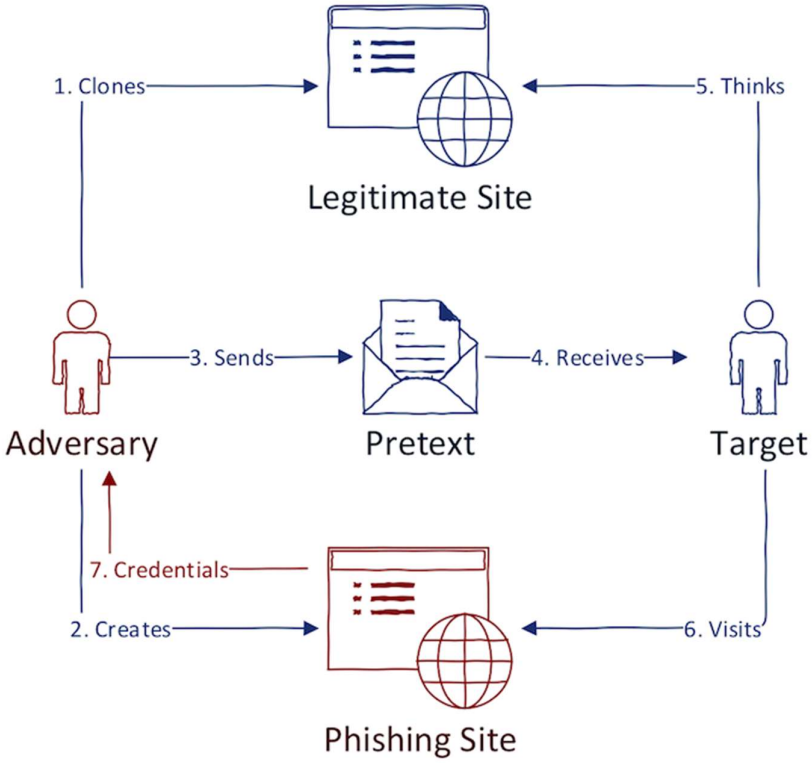

Image 1: Phishing Work Flow Chart

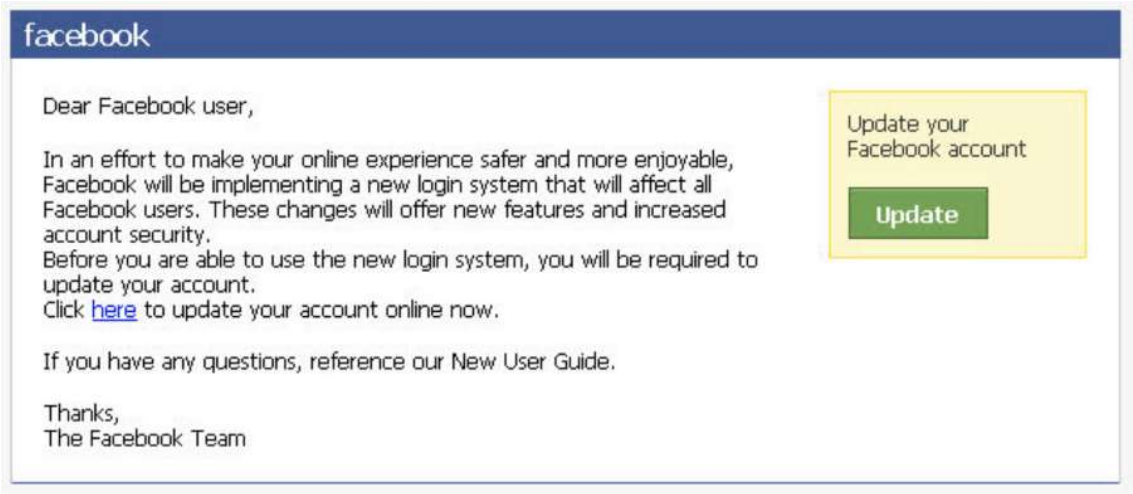

Image 2: Example of Spoof Message

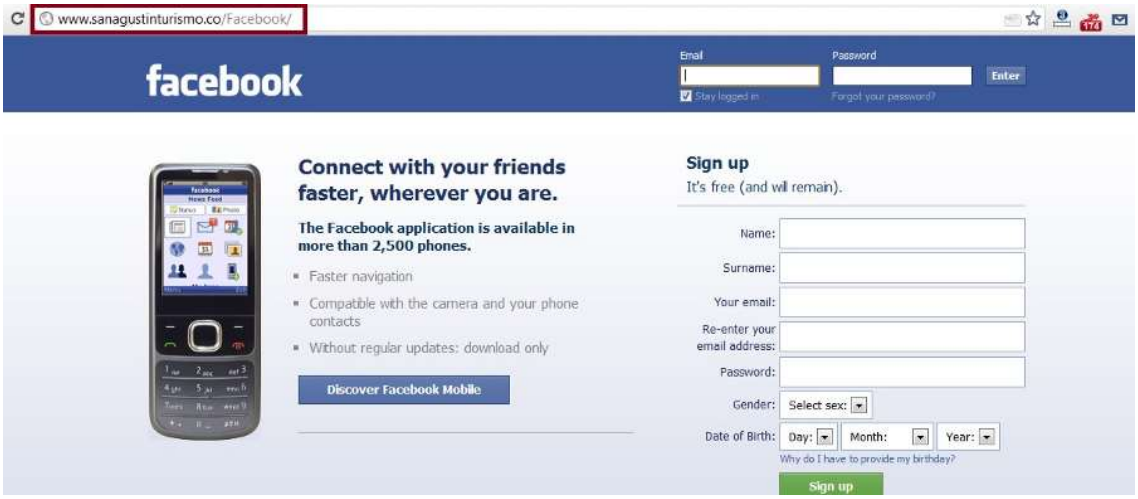

Image 3: Fake URL of Phishing Website when you Click it.

- **Necessities for Project:**

  - (Anaconda for Jupiter Notebook and Python) or you can use (Google Colab)
  - Python libraries: (Pandas and Numpy)
  - Natural Language Processing Libraries like (Countvectorizer and Tfidfvectorizer)
  - Data Set https://github.com/Jcharis/Machine-Learning-In-Julia-JCharisTech/blob/master/urldata.csv

- **Dataset Description:**

  Dataset consist of Two Columns: (url, label) and 420465 rows.

  url – is list of URLs'
  Label – Consists of (good, bad) labels for URLs'

- **How Project Works:**

  1. Import all Python libraries needed (Pandas and Numpy).
  2. Import all NLP libraries needed (Countvectorizer, Tfidfvectorizer).
  3. Import your dataset.

4. **Use of Natural Language processing (NLP):**

    - NLP is used to extract features from the URL provided. From NLP we first tokenize the given URL, means we first break our URL into pieces of tokens or chunks.

    - This is done by makeToken() function that we have defined. We call this function in vectorizer Tfidfvectorizer() function which in turn converts the tokens in sparse matrix based on presence of tokens in URL. 1 for present and 0 for not.

    - Now this sparse matrix becomes our feature dataset.

5. Now we split the dataset in Training and Testing data. And afterwards, since it's a classification problem we use linear regression to train it.

6. We make this classification on the basis of presence of the tokens in URLs' in our dataset.

7. Calculate Accuracy and Collect input for predictions.

- **<u>Applications:</u>**

    1. This Project can be used to detect any malicious link that might be in your Email inbox or anywhere else.

    2. Phishing is a Criminal offence and strong detection techniques must be developed to resolve these issues, since this act can not only steal money from your bank account, it can devastate you on emotional level also by stalking you or hampering your valuable secrets so projects like this are very necessary.

- **<u>Acknowledgement:</u>**

    We would like to acknowledge my mentor **Mr. Bipul Shahi Sir** for mentoring in this field of machine learning. We really thank you Sir wholeheartedly.

    Because of you we were able to make this awesome project.