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 of 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.

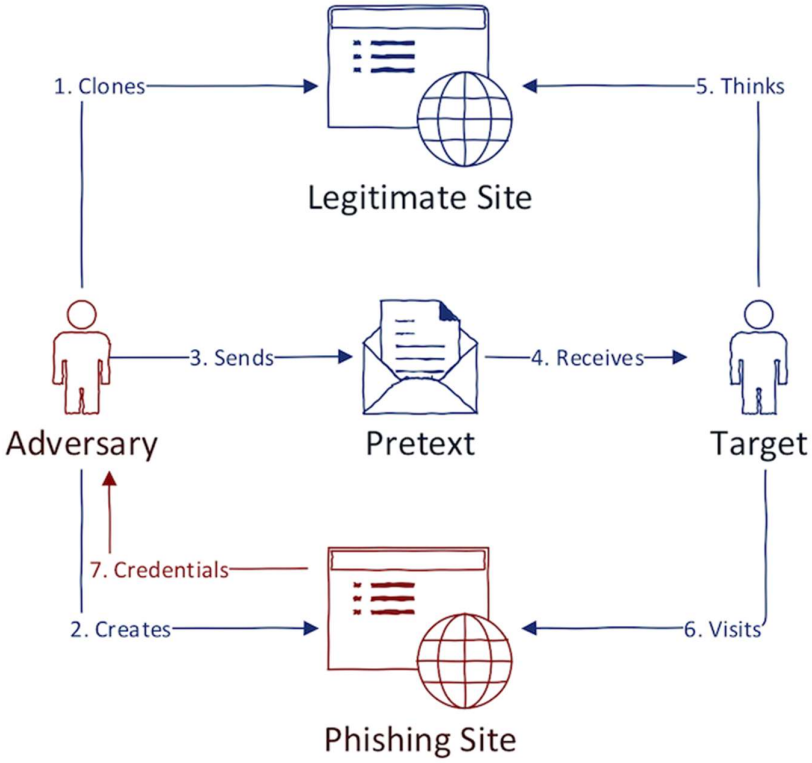

Image 1: Phishing Work Flow Chart

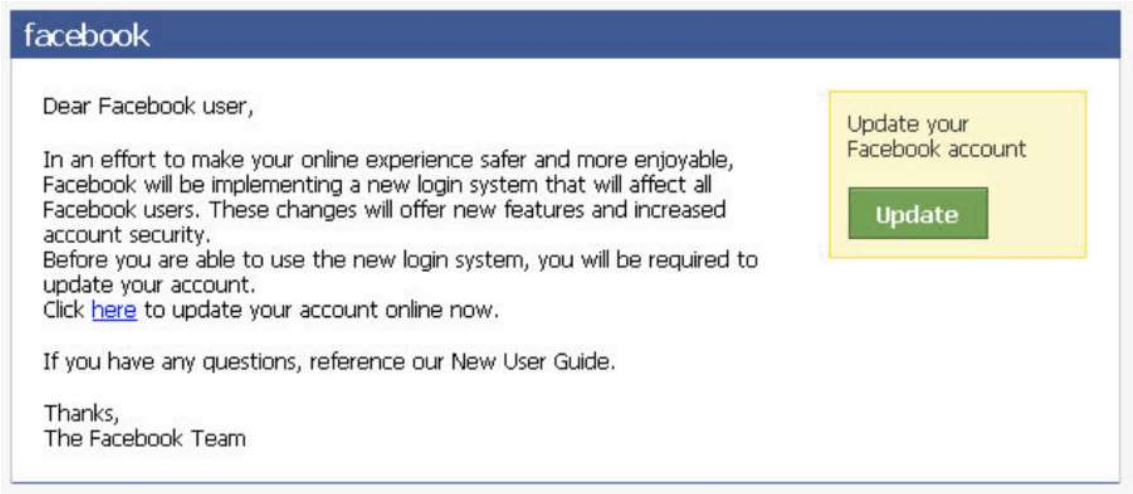

Image 2: Example of Spoof Message

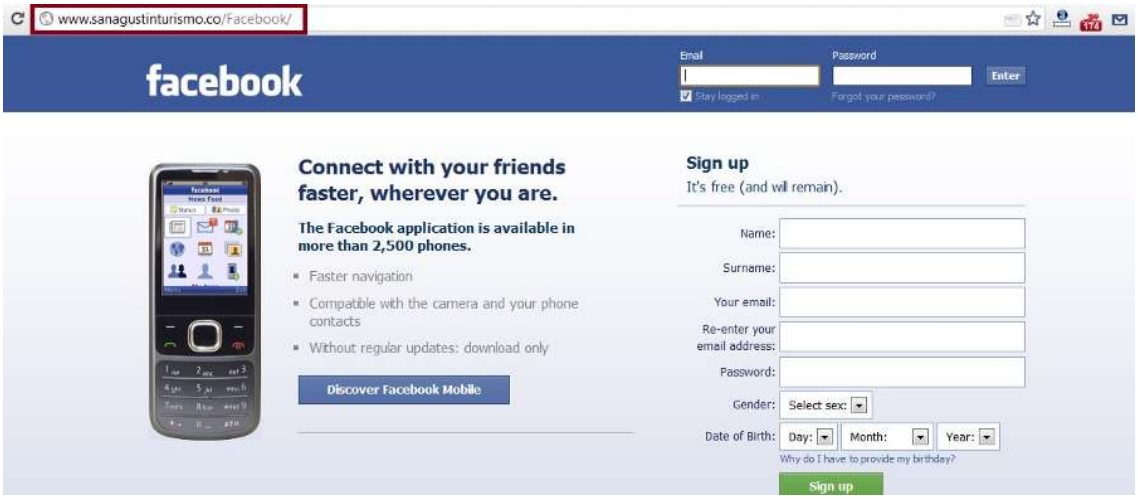

Image 3: Fake URL of Phishing Website when you Click it.

- **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 NLP for features extraction from url as features.

5. Using Tfidfvectorizer create token vectors.
6. Create sparse matrix of these.
7. Now Separate Features and labels
8. The 'Label' field will be label.
9. The Sparse matrix will be features.
10. Split and train the data using Logistic Regression. We have used logistic regression you can use any other models like decision trees or naïve bays etc.
11. Calculate Accuracy and Collect input for predictions.

- **Applications:**

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.

- **Acknowledgement:**

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.