8 - Fast Webpage Classification with Robust Phishing Detection

Soeren Hougaard Mulvad, Kanav Sabharwal, Liu Xiaotong, Niu Yibo

{shmulvad,e0575775,e0686132,e0691793}@u.nus.edu

Introduction

With the vast growth of the internet, effective webpage classification has become ever more essential. However, typical classification requires every page to be fetched and analyzed which is slow. We propose **DURC**, a **d**eep **UR**L **c**lassifier that can *identify the general classes* while being able to *detect phishing webpages* using *only the URL*.

Research questions

The questions we would like to answer are:

- How do we solve the overall problem of doing website classification with only the URL?
- How do different parameters, features, and architectures affect the model performance and what achieves the best performance?
- How can we achieve robust phishing detection?

Datasets

We employ DMOZ and ILP for general classification while PHISHING is used for phishing detection.

	Rows	Classes	Categories
DMOZ	1,500k	15	General
ILP	8k	7	Academia
PHISHING	45k	2	Phishing

GitHub Repo

Hand-picked Features

We extract **20 features** in total, including the number of characters, digits, capitals, and words in each part of the URL (domain, path, arguments). Furthermore, we employ features useful for detecting phishing websites, such as whether an IP address is present in the hostname, security-sensitive words, and if the URL contains a suspicious symbol, such as @-symbol, dots in path and arguments, etc. [1].

Word and Character Embeddings

The pre-trained (P) word embeddings WORD2VEC and GLOVE were initially used for feature engineering. But these have shortcomings for URLs. We switched to self-trained (S) FASTTEXT embeddings [2] that also handle character-level embeddings better. On ILP with random forest, the F1 scores are:

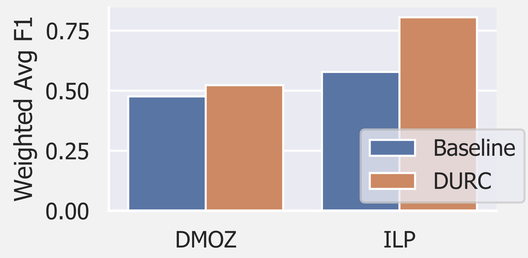
(P) WORD2VEC	(P) GLOVE	(S) FASTTEXT
0.6839	0.7104	0.7399

General Classification Results

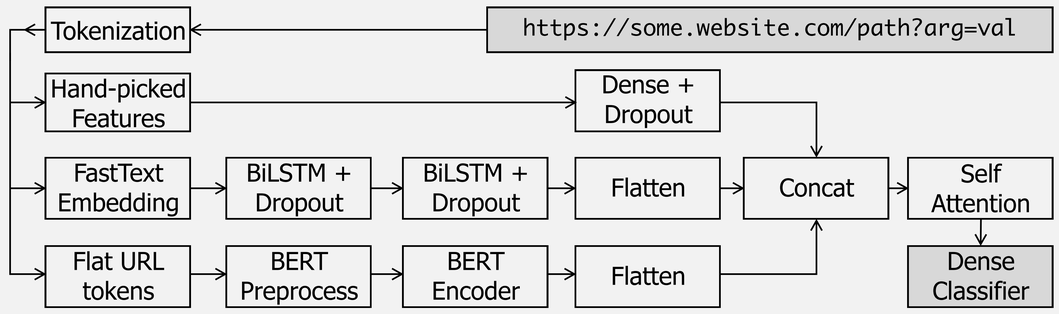
It was explored whether to use LSTM, BiLSTM, or CNN for the embedding input and whether to use self-attention or dropout at the end. The weighted avg F1 scores on DMOZ are shown.

	SELFATTENTION	DROPOUT
LSTM	0.2858	0.4872
BILSTM	**0.5060**	0.4926
CNN	0.4798	0.4948

With BiLSTM and self-attention determined to be superior, we tuned the hyperparameters. Our model is compared with a baseline that is constructed by feeding the flat URL tokens into a TfIdfVectoriser and then running logistic regression on the vectors:



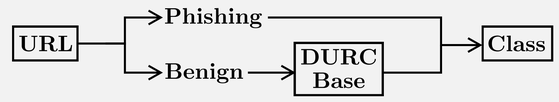
DURC-Base: General Classification Model Architecture



DURC-Base: The URL is tokenized from which a hand-picked feature vector, FastText embedding matrix and flat tokens are created. They are fed into a dense model, BiLSTM model and BERT. The outputs are concatenated, fed to a self-attention layer and forwarded to the final dense classifier.
NB: Only the general classification part, DURC-Base, is shown in figure.

Robust Phishing Dectection

For robust phishing detection [3], we employ a separate model trained specifically with this task in mind. The URL is fed to the phishing classifier. If deemed benign, it is forwarded to DURC-Base and otherwise phishing is the returned output. The DURC flow is:



High accuracies can easily be achieved on the phishing dataset, so we compare our phishing classifier to a similar baseline as before on the error rate (lower is better).

BASELINE	DURC
0.9817%	0.2805%

Conclusion

We have created a deep URL classifier, DURC, that classifies webpages while being able to detect phishing webpages using only the URL. The model employs several novel techniques for an accurate classification. The model is superior to the baseline with a 57.81% → 80.57% avg macro F1 on ILP while having only a 0.28% phishing error rate.

References

[1] Min-Yen Kan and Hoang Thi.
Fast webpage classification using url features.
2005.

[2] Piotr et. al. Bojanowski.
Enriching word vectors with subword information.
2017.

[3] Rakesh Verma and Avisha Das.
Fast feature extraction and malicious url detection.
2017.