**SMS phishing detection based on “Normal” traffic learning**

**Abstract**

In recent years, spammers got interested in SMS as a method of attacking potential targets. Spam SMS can trick people into giving up their confidential information, which may lead to severe violations of their privacy. SMS phishing has become a threat to many users and thus requires a precise solution. Machine learning algorithms have emerged as a great tool to classify data into labels. One of the problems with SMS phishing is the lack of labeled data that can be exploited while researching and discovering patterns that define ham messages. In this article, we will investigate the approach of machine learning which will automate the way of detecting false messages and prevent them from ever reaching their destination, but also investigate an approach that will allow us to automatically increase the dataset.

**Introduction**

Mobile messaging is one of the most popular ways of communicating today, having billions of mobile devices exchange numerous messages. However, this method of communicating is very insecure due to the lack of a proper message filtering mechanism. One cause of such insecurity is spam, and it makes mobile message communication insecure. Spam is any kind of unwanted, unsolicited digital communication that gets sent out in bulk. Often spam is sent via email, but it can also be distributed via text messages, phone calls, or social media. It contains different forms such as adult content, selling items or services, and so on.

Ways of overcoming these threats are offered in this article. Many propose a method based on machine learning classifiers to classify ham or spam. This will be indeed one of the approaches we will try. However, we won’t be using a simple classification algorithm such LR or KNN.

**DistilBERT**

Bert - Bidirectional Encoder Representations is a transformer-based machine learning technique for natural language processing pre-training developed by Google. BERT can be used for a wide variety of language tasks, while only adding a small layer to the core model: Classification tasks such as sentiment analysis are done by adding a classification layer on top of the Transformer output for the [CLS] token.

DistilBERT is a small, fast, and cheap Transformer model based on the BERT architecture. The model uses a technique called distillation, which approximates Google's BERT, i.e. the large neural network by a smaller one. The idea is that once a large neural network has been trained, its full output distributions can be approximated using a smaller network.

Transformers are python based libraries. In this attempt, we use a pre-trained Natural Language Understanding (NLU) model - from Hugging Face, and fine-tune it locally to best suit our dataset.

The process of fine-tuning DistilBERT is comprised of similar steps done in other machine-learning models. Firstly, we normalize the labels by changing them from ‘ham/spam’ to 0 and 1. Secondly, we split the data into train and test batches, after examining that the data is balanced. We then call the BertTokenizerFast to run end-to-end tokenization on the dataset: punctuation splitting and word pieces. The tokenizer also converts the data into numerical vectors that the model can process. Finally, we can train the fine-tuned model and export it for test use.

The DistilBERT had a 100% accuracy on the test batch and was able to predict correctly each message being ham or spam.