**Malicious PDF File Detection - Commitment Document**

*Alexey Titov, Shir Bentabou*

*Supervisors: Dr. Amit Dvir, Dr. Ran Dubin*

**Accomplished Tasks:**

* Hand in project abstract.
* Hand in project proposal to supervisors.
* Hand in project poster.
* Deeply know PDF file structure, features and fields.
* Study about phishing, URLs, and JavaScript uses in PDF files.
* Research methods from previous researches.
* Build the plan and schedule for our project: phases, tasks in each phase, deadlines for each task.
* Research existing tools for our usage in the project.
* First phase – Researching and creating our work tools:
  + Extracting telemetry.
  + Extracting text from picture.
  + Extracting text from pdf file (using PDFMiner).
  + Extracting URLs (using pyPDF).
  + Extracting URLs from JS in the file tags (using peePDF).
  + Extracting preview of a PDF file (using PIL + pdf2image).
* Second phase – Creating an image-based classification machine:
  + Research vector features.
  + Building the feature vector.
  + Applying machine learning algorithms on the feature vector.
* Third phase – Creating a text-based classification machine:
  + Research vector building methods.
  + Applying the vector methods on our samples.
  + Applying machine learning algorithms on the text vector.
  + Applying a deep learning method on the text vector.
* Fourth phase – Creating a classification machine based on PDF tags, JS, URLs, objects and streams:
  + Researching the features that will build the vector for this machine in each one of the four parts: PDF tags, JS, URLs, objects and streams.
  + Research existing tools for the extraction of the features chosen (JAST, Analyze PDF, peePDF).
  + Extraction of the features from samples.
  + Building the feature vector.
  + Applying machine learning algorithms on the feature vector.
  + Applying a deep learning method on the feature vector.

**To Be Accomplished:**

* Prepare project day presentation.
* Writing project book.
* Fifth phase – Creating an ensemble machine:
  + Combining the three machines into an ensemble machine.
  + Determining the overall classification method for the ensemble machine.
  + Applying machine learning algorithms to ensemble machine:
    - Random Forest, AdaBoost (Adaptive Boosting), Gradient Tree Boosting, XGBoost.
* Improvement phase - Deciding improvement phase aim, and numeric success rate for each classifier, and ensemble machine as well.
* Improvements for each phase:
  + Second phase:
    - Try to improve picture classification in two ways:
      * Applying additional vector building methods (such as near similar image matching).
      * Applying additional machine learning algorithms on the vectors.
  + Third phase:
    - Try to improve text classification in the following way:
      * Applying additional machine learning algorithms on the different vector building methods (word2vec, TF-IDF) to achieve better results.
  + Fourth phase:
    - Improve feature selection in the following ways:
      * Random choice method.
      * Summing features method.
      * Combining features as new features in vector.
    - Applying additional machine learning algorithms on the vectors.
* Overall improvements:
  + Applying iterative retraining methods on the machines.