Review of Bachelor's Thesis

Student: Holop Patrik
Title: Classification of Potentially Malicious File Clusters via Machine Learning (id 21927)
Reviewer: Zendulka Jaroslav, doc. Ing., CSc., UIFS FIT VUT

1. **Assignment complexity** — **considerably demanding assignment**
   Obtížnost zadání spočívala jednak v použití netradičního přístupu ke klasifikaci potenciálních hrozeb na základě společných vlastností shluků souborů, jednak v očekávaném rozsáhlém experimentování.
2. **Completeness of assignment requirements** — **assignment fulfilled with enhancements**
   Zadání již bylo koncipované jako značně náročné, výrazně přesahující průměrně obtížná zadání. Všechny body považuji za kvalitně splněné.
3. **Length of technical report** — **in usual extent**
   Práce je poměrně rozsáhlá, rozsah je na horní hranici běžného rozmezí. Veškerý obsah považuji za relevantní.
4. **Presentation level of technical report** — **99 p. (A)**
   Práce má logickou strukturu a kapitoly na sebe navazují. Kapitoly jsou informačně velice bohaté. Mohu konstatovat, že obsah práce jasně ukazuje, jaké množství kvalitní práce student při zpracování zadání odvedl. Velice kladně hodnotím jak část rešeršní, tak experimentální a realizační.
   Obsah je v naprosté většině dobře srozumitelný a pochopitelný. V některých případech by prospělo doplnit text ještě názorným obrázkem. Týká se to především kapitoly 3.1 k systému Clusty, na který systém Hamlet vyvíjený v práci navazuje. Obrázek, který by ukazoval základní kroky zpracování systémem Clusty (podobně jako později obr. 7.2 u systému Hamlet) a současně naznačil návaznost části vytvářené v rámci bakalářské práce, by přispěl k lepšímu náhledu na celkový kontext. Toto ale není připomínka, která by nějak snižovala velmi pozitivní dojem, který z práce a dosažených výsledků mám.
5. **Formal aspects of technical report** — **100 p. (A)**
   Po typografické stránce nemám připomínek, práce je kvalitní. Práce je psaná velmi dobrou angličtinou ve které se jenom zřídka vyskytují drobné chyby.
6. **Literature usage** — **100 p. (A)**
   Práci s literaturou hodnotím velmi pozitivně. Seznam použité literatury zahrnuje celkem 40 zdrojů, což je počet výrazně převyšující počty běžné u bakalářských prací. Nejedná se zdaleka jen o internetové zdroje, nýbrž i sborníky z konferencí a knižní publikace. Zdroje jsou v textu řádně citovány.
7. **Implementation results** — **100 p. (A)**
   Realizační výstup, do kterého řadím i experimentální část, považuji za velmi kvalitní. Experimenty byly zaměřeny na přípravu datových souborů a jejich použití pro vytvoření zvolených klasifikačních modelů, na základě nichž byly potom vybrány klasifikační metody a soubory, u kterých bylo dosaženo nejlepších výsledků. Ty potom byly použity v systému/službě Hamlet pro klasifikaci shluků. Popis a diskuse výsledků experimentů i návrh a implementace systému Hamlet svědčí o výborných schopnostech a dovednostech studenta.
8. **Utilizability of results**
   Jde o práci rozšiřující již publikované výsledky, ale i přinášející některé nové poznatky, zejména z výsledků provedených experimentů. Systém Hamlet byl vytvořen pro interní využití ve společnosti Avast, kde se také podle informací v závěru práce používá.
9. **Questions for defence**
   1. Na základě čeho byly voleny hodnoty hyperparametrů u metody Random forest?
   2. V závěru uvádíte, že z experimentů vyplynulo, že klasifikace malware na úrovni shluků souborů typu PE, APK a .NET je možná s nižší přesností než při klasickém použití úrovně souborů. O jak velký rozdíl se jedná a jakou výhodou je vyvážen?
10. **Total assessment** — **100 p. excellent (A)**
    Jde o velmi kvalitní, nadstandardní práci splňující v plném rozsahu náročné zadání. Velice kladně hodnotím jak část rešeršní, tak experimentální a realizační. Podrobnosti jsou uvedeny výše v dílčích hodnoceních.

In Brno 30. May 2019

.....
signature