Posudek oponenta bakalářské práce

Student: Setinský Jiří
Téma: Detekce škodlivých doménových jmen (id 23737)
Oponent: Perešíni Martin, Ing., UITS FIT VUT

1. Náročnost zadání **průměrně obtížné zadání**

Práce má **průměrně obtížné** zadání, student si musel prostudovat různé techniky strojového učení a využití strojového učení v oblasti detekce doménových jmen DAG v síťovém prostředí. Zadání bylo sice středně obtížné, ale student k němu přistoupil zodpovědně a zpracoval ho kvalitně.

2. Splnění požadavků zadání **zadání splněno**

Student **splnil všechny** body zadání. Pozitivní také je, že student částečně **rozšířil** své řešení o predikci úspěšnosti nově vytvořených datových sad pro porovnání klasifikátorů. Studentovo řešení je porovnáno s existujícími řešeními, ale z mého pohledu nedostatečně, neboť není jasné, zda byla řešení porovnávána na vhodných datech. Ve výsledku práce působí dojmem, jako by vytvořené řešení bylo "state of the art" v dané oblasti, což se mi osobně nepozdává (může to být způsobeno právě výběrem nesprávné datové sady).

3. Rozsah technické zprávy **přesahuje obvyklé rozmezí**

Rozsah práce je na **horní hranici**, počet normostran se blíží k 90. Technická zpráva kromě informací relevantních k zadání obsahuje i části které jsou nadbytečné nebo redundantní.

4. Prezentační úroveň předložené práce **72 b. (C)**

Práce jako celek má určitou logickou strukturu, ale některé kapitoly obsahují logické nesrovnalosti nebo nenavazují na další části textu a tím narušují čtení textu (celkový tok je přerušen). Příkladem je kapitola 3, která popisuje doménová jména a pak najednou kapitola popisuje systém NEMEA nebo software rapidminer, který patří do strojového učení. Úroveň prezentace určitých obrázků je mizivá - na obrázku 4.4 nejsou popsány osy a není jasné, co vyjadřuje histogram na diagonále nebo jak autor myslel lineární závislost prezentovaných dat. Totéž platí pro obrázek 7.4. Kapitola 7 by měla být přepracována a začleněna do textu jiným způsobem a v neposlední řadě se mi nelíbí, že student retrospektivně popisuje své řešení. (Nejprve bylo představeno řešení a poté jak se k němu dospělo. To by bylo v pořádku, ale některé souvislosti jsou vysvětleny v textu až později.) Celkově je úroveň prezentace **průměrná**.

5. Formální úprava technické zprávy **77 b. (C)**

Jelikož je práce psána v českém jazyce a já nejsem rodilý mluvčí, nemohu se vyjádřit k jazykové správnosti a gramatičnosti práce, přesto bych měl několik připomínek. V textu se často vyskytuje popis v trpném rodě budoucího času ("kapitola bude popisovat"), nahradil bych ho časem přítomným ("v kapitole je popsáno"). Z typografického hlediska má práce určité **nedostatky**. Na straně 16 je uveden přehled některých algoritmů DGA, tento přehled poměrně dosti splývá s okolním textem a navrhoval bych jej buď přehledně umístit do tabulky, nebo alespoň přesunout do prostředí *itemize*. Dále bych navrhoval používat jednotnější styl, nepoužívat tak často nečíslované podkapitoly, upravit a odstranit nepříjemné mezery v textu (str. 53), někdy je vhodné použít kurzívu na zvýraznení, nepoužívat nekvalitní rastrové obrázky, apod.

6. Práce s literaturou **88 b. (B)**

Student využívá relevantní zdroje v dostatečném množství, čerpá informace z odborných textů a dokumentace, ale také z webových stránek a příruček dostupných především na internetu. Problém vidím zejména ve dvou citacích [18,26], které odkazují na Stack Overflow a Wikipedii. Takové citace by se v odborném textu vyskytovat neměly, přesto hodnotím práci s literaturou jako **uspokojivou**.

7. Realizační výstup **90 b. (A)**

Realizační výstup je vyhovující a odpovídá specifikaci. Výstupem implementace je vytvoření klasifikátoru škodlivých doménových jmen. Kromě samotného klasifikátoru DGA adres se studentovi podařilo vytvořit také nový klasifikátor pro predikci úspěšnosti různých datových sad.

8. Využitelnost výsledků

Práce byla vyvinuta ve spolupráci s organizací CESNET a její implementace je integrována jako modul systému NEMEA, který slouží k monitorování síťového provozu. Výsledky byly testovány na reálných provozních datech a zdá se, že klasifikátor je připraven k nasazení v praxi. Výsledky práce budou pravděpodobně dále využity v systému NEMEA a existuje také možnost rozšiřování a zdokonalování klasifikátoru.

9. Otázky k obhajobě

1. Ve své práci jste se zmínil o použití nástrojů *host* a *whois* pro rezoluci DNS jmen. Zmínil jste také, že použití

těchto nástrojů je příliš pomalé, a proto nevhodné. Uvažoval jste o použití těchto nástrojů paralelním způsobem?

2. Jaké vzorky DGA (různý malware používá různé generátory) jste použil ve svých souborech dat? Použil jste vzorky pouze jednoho typu, nebo mix řekněme 100 různých? V teoretické části uvádíte například Bambenek nebo Netlab 360, ale v praktické části jsem si nevšiml, jaké konkrétní DGA vzory jste použil.

10. Souhrnné hodnocení **82 b. velmi dobře (B)**

Student splnil všechny povinné body zadání. Praktická část práce je zpracována vhodně, až na některé nedostatky týkající se výběru datové sady. Kromě toho vyhotovení obsahuje rozšíření mimo samotné zadání. Textová část práce je z hlediska kvality a provedení horší. Výsledky jsou použitelné v rámci systému NEMEA, což hodnotím kladně. Celkově hodnotím práci jako nadprůměrnou a studentovi navrhuji **B**.

Prohlášení: Uděluji VUT v Brně souhlas ke zveřejnění tohoto posudku v listinné i elektronické formě.

V Brně dne: 3. června 2021

Perešíni Martin, Ing.
oponent