МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ
ФАКУЛЬТЕТ КОМП'ЮТЕРНО-ІНФОРМАЦІЙНИХ СИСТЕМ І ПРОГРАМНОЇ ІНЖЕНЕРІЇ

СТЕФАНІВ АНДРІЙ МИХАЙЛОВИЧ

УДК 004.491

МЕТОДИ ОБРОБКИ ПРИРОДНЬОЇ МОВИ ІЗ ВИКОРИСТАННЯМ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ SPARK MLLIB

122 «Комп'ютерні науки та інформаційні технології»

Автореферат
дипломної роботи на здобуття
освітнього рівня «магістр»

Тернопіль
2018

Роботу виконано на кафедрі комп'ютерних наук Тернопільського національного технічного університету імені Івана Пулюя Міністерства освіти і науки України

Керівник роботи: кандидат технічних наук, В.О. завідувача кафедри кібербезпеки
Загородна Наталія Володимирівна,
Тернопільський національний технічний університет імені Івана Пулюя,

Рецензент: кандидат технічних наук, доцент кафедри кібербезпеки
Боднарчук Ігор Орестович,
Тернопільський національний технічний університет імені Івана Пулюя,

Захист відбудеться 21 лютого 2018 р. о $10^{.00}$ годині на засіданні екзаменаційної комісії №__ у Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56, навчальний корпус №1, ауд. 70

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми роботи. Поява мережі Інтернет та бурхливе зростання доступної текстової інформації значно прискорило розвиток наукової галузі, яка існує вже багато десятків років і відомої як автоматична обробка текстів (Natural Language Processing) та комп'ютерна лінгвістика (Computational Linguistics). В рамках цієї галузі запропоновано багато перспективних ідей з автоматичної обробки текстів на природній мові, які були втілені в багатьох прикладних системах, в тому числі комерційних. Сфера додатків комп'ютерної лінгвістики постійно розширюється, з'являються все нові завдання, які успішно вирішуються, в тому числі із залученням результатів суміжних наукових областей. У сучасному суспільстві для задоволення його потреб виникають проблеми інформаційного забезпечення всіх сфер діяльності людини. Одна з таких проблем – забезпечення надійного захисту приватної інформації. Особливої гостроти вона набуває у зв'язку з масовою комп'ютеризацією всіх видів діяльності людини, при об'єднанні ЕОМ у комп'ютерні мережі та підключення до Internet. Зростає кількість атак, мета яких зробити веб-користувачів впевненими в тому, що вони спілкуються з довіреним юридичним особам з метою крадіжки інформації про обліковий запис, облікових даних для входу в систему та інформації про ідентичність в цілому. Цей метод нападу, широко відомий як "фішинг", найчастіше ініційований шляхом надсилання електронних листів із посиланнями на підроблені веб-сайти, які збирають конфіденційну інформацію користувачів. Тому постає необхідність у дослідженні і розробці нових оптимальних підходів до виявлення фішингових електронних листів.

Мета роботи: розробка і дослідження нового підходу до визначення фішингових електронних листів на основі обробки природньої мови із використанням інформаційної технології Spark MLlib.

Об'єкт, методи та джерела дослідження. Основним об'єктом дослідження є процес обробки природньої мови через розробку моделі класифікації електронних листів на фішингові та нейтральні.

Наукова новизна отриманих результатів:

- вдосконалено метод класифікації електронних листів на такі два типи як фішингові та нейтральні;
- представлено реалізацію даного підходу із використання інформаційної технології Spark MLlib.

Практичне значення отриманих результатів.

В результаті дослідження розроблено новий підхід у визначенні фішингових електронних листів із використання інформаційної технології Spark MLlib.

Апробація. Окремі результати роботи доповідались на VI Міжнародної науково-технічної конференції молодих учених та студентів «Актуальні питання сучасних технологій». Тернопіль, ТНТУ, 16-17 листопада 2017 року та на VII Університетській конференції студентів та молодих вчених "Інженер XXI століття". Бельско-Бяла, Університет Бельско-Бяла, 8 грудня 2017 року.

Структура роботи. Робота складається з розрахунково-пояснювальної записки та графічної частини. Розрахунково-пояснювальна записка складається з вступу, 7

частин, висновків, переліку посилань та додатків. Обсяг роботи: розрахунково-пояснювальна записка – арк. формату А4, графічна частина – 7 аркушів формату А1

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі проведено огляд сучасної проблеми визначення фішингових електронних листів та охарактеризовано основні завдання, які необхідно вирішити.

В першій частині проведено аналіз стану питання за літературними та іншими джерелами. Здійснено аналіз предметної області. Дано визначення поняттю обробка природньої мови. Розглянуто походження та підрозділи обробки природньої мови.

В другій частині наведено опис технологій, які використовувались при дослідженні. Розглянуто проблему фішингу. Описано традиційні методи їх вирішення. Описано алгоритми класифікаторів та набори вхідних даних.

В третій частині наведено розробку комплексної моделі класифікації. Описано процес попередньої обробки вхідних даних. Наведено основні характеристики, які використовувались для класифікації. Здійснено навчання і тестування кожного класифікатора і комплексної моделі. Проведено аналіз отриманих метрик.

В спеціальній частині описано кластерний режим Apache Spark. Розглянуто архітектуру Spark-додатку. Наведено основі кроки запуску Spark-додатку в кластерному режимі.

В частині «Обґрунтування економічної ефективності» проведено розрахунки техніко-економічної ефективності впровадження розрахунку ефективності тестування (або впровадження дослідження розрахунку ефективності тестування).

В частині «Екологія» розглянуто методи узагальнення екологічної інформації та вимоги до приміщень для експлуатації моніторів і ПОЕМ. Наведено шляхи дотримання цих вимог.

В частині «Охорона праці та безпека в надзвичайних ситуаціях» розглянуто атестацію робочих місць користувачів ЕОМ. Наведено гігієнічну оцінку умов праці програміста. Описано фактори, що впливають на функціональний стан користувача комп'ютера. Досліджено інженерний захист персоналу об'єкту та населення і правила їх застосування.

У загальних висновках щодо дипломної роботи описано основні результати, які були досягнуті в результаті дослідження.

В додатках до пояснювальної записки приведено тези та лістинг програмної реалізації розробленого підходу.

ВИСНОВКИ

Досліджено методи визначення фішингових електронних листів із використанням одного з класифікаторів: логістичної регресії, дерев прийняття рішень та методу опорних векторів. Розроблено і досліджено підхід із поєднанням трьох класифікаторів. Проаналізовано та порівняно точність, прецизійність і повноту нового підходу з іншими моделями.

СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

1. Долінський Т. Методи визначення фішингу із застосуванням технології Apache Spark MLlib [Текст] / Долінський Т.М. Стефанів А.М., Козак Р.О.. Тези доповіді VI Міжнародної науково-технічної конференція молодих учених та студентів "Актуальні задачі сучасних технологій". – Тернопіль, ТНТУ, 2017. – с. 224.
2. Dolinskii T. M. Using machine learning algorithms of apache spark mllib for detection of phishing in text data [Text] / Dolinskii T. M., Stefaniv A.M., Kozak R. O.. Paper from the VII Inter University Conference ofStudents and Young Scientists "Engineer of XXI Century". – Bielsko-Biala, University of Bielsko-Biala, 2017. – с. 414.

АНОТАЦІЯ

Дипломна робота присвячена дослідженню методів обробки природньої мови із використанням інформаційної технології Spark MLlib.

Метою роботи є розробка і дослідження нового підходу до визначення фішингових електронних листів на основі обробки природньої мови із використанням інформаційної технології Spark MLlib.

Об'єктом дослідження є процес обробки природньої мови через розробку моделі класифікації електронних листів на фішингові та нейтральні.

Предметом дослідження є методи обробки природньої мови, класифікатори та інформаційна технологія Spark MLlib, які можуть бути запроваджені задля забезпечення максимальної ефективності визначення фішингових електронних листів.

Основні результати: досліджено існуючі методи класифікації електронних листів, проведено тестування і визначення оптимальності використання уже існуючих методів, розроблено та досліджено новий підхід у класифікації фішингових електронних листів із використанням комплексної моделі на основі трьох класифікаторів, досліджено коректність роботи даної моделі, порівняно результати із іншими методами.

Ключові слова: ОБРОБКА ПРИРОДНЬОЇ МОВИ, КЛАСИФІКАЦІЯ ТЕКТОВИХ ДАНИХ, СЕМАНТИЧНИЙ АНАЛІЗ, ФІШИНГ, SCALA, SPARK, МАШИННЕ НАВЧАННЯ.

ANNOTATION

The graduate work is devoted to research of methods of natural language processing using Spark MLlib information technology.

The purpose of the work is to develop and investigate a new approach to the definition of phishing emails based on the processing of natural language using the information technology Spark MLlib.

The object of the study is the process of processing natural language through the development of a model for the classification of emails for phishing and neutral.

The subject of the study is the natural language processing methods, classifiers and Spark MLlib information technology that can be implemented to maximize the effectiveness of phishing emails.

Main results: the existing methods of classification of emails were investigated, the testing and determination of the optimality of the use of existing methods was conducted, a

new approach in the classification of phishing emails was developed and investigated using a complex model based on three classifiers, the correctness of the work of this model was investigated, the results compared with other methods .

Scientific novelty of the development is improvement of the method of classifying emails for two types as phishing or neutral and the implementation of this approach with the use of Spark MLlib is presented.

Key words: NATURAL LANGUAGE PROCESSING, CLASSIFICATION OF TEXT DATA, SEMANTIC ANALYSIS, FISHING, SCALA, SPARK, MACHINE LEARNING.