УДК 004.056:004.912

ЗАСТОСУВАННЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ТЕКСТІВ ДЛЯ ВИРІШЕННЯ ЗАДАЧ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ. ЧАСТИНА 2. ПОШУК ІНСАЙДЕРІВ

Панков Д.С., Машура А.П.
e-mail: andrii.mashura@nure.ua
Науковий керівник - к.т.н., доц. Горелов Д.Ю.
Харківський національний університет радіоелектроніки, каф. КРіСТЗІ
студ. наук. гурток «Біометричні технології контролю доступу»
м. Харків, Україна

An algorithm for detecting insiders based on automated semantic analysis of sent emails is proposed. In the Orange software environment, based on the database "The Enron Email Dataset", a study of the possibilities of classifying the organization's personnel into insiders and ordinary employees was carried out. The accuracy of identifying insiders was 60%.

Витік даних є однією з найнебезпечніших загроз для сучасних компаній. Для запобігання витокам традиційно застосовуються системи класу DLP, що дають змогу виявляти конфіденційну інформацію в потоках даних, які залишають інформаційний периметр організації. Однак дедалі більше експертів сходиться на думці, що використання DLP-систем недостатньо ефективне, і тому витоки необхідно визначати ще до стадії пересилання даних за інформаційний периметр. Це твердження ґрунтується на дослідженнях, які показують, що від моменту, коли користувач вирішує вкрасти дані до безпосередньо пересилання даних, проходить від кількох тижнів до кількох місяців, які йдуть на стадію підготовки витоку. У цій стадії поведінка користувача відрізняється від його звичайної легітимної активності як за набором виконуваних дій, так і за змістом оброблюваної інформації.

В останнє десятиріччя з’явився ряд робіт, в яких досліджуються можливості використання технологій інтелектуального аналізу текстів для пошуку інсайдерів [1-3]. Відповідно до цих робіт кожному з типів інсайдерів властиві унікальні фрази, за якими їх можна ідентифікувати. В роботі [4] наведено результати класифікації співробітників організації на 5 типів (4 класи інсайдерів та звичайний співробітник). Заявлена точність – 92 %.

Для проведення експериментальних досліджень в якості тестового датасету було обрано «The Enron Email Dataset». Це база даних із понад 600000 електронних листів, створених 158 співробітниками корпорації Enron за роки, що передували банкрутству компанії в грудні 2001 року. База даних була створена із серверів електронних листів Enron Федеральною комісією з регулювання енергетики під час розслідування. Згодом копія бази даних електронних листів стала доступна дослідникам.

З датасету було відібрано 5 співробітників з тегом «POI-true» (тобто відомо, що це зловмисник) та 5 співробітників з тегом «POI-false». Для кожного з них з датасету були виокремлені теки з тегом «sent» та прибрані з аналізу теки з тегами «received», «deleted», «inbox», оскільки, на нашу думку, саме відправлені листи мають містити в собі ознаками злочинної діяльності співробітника, бо тут можуть міститись алгоритми дій для сторонніх зловмисників, інформація про бонуси, отримані злочинними діями, спонукання до злочинної співпраці інших співробітників, особистісна інформація, що має ознаки злочинних дій (кредити, трати, оренди тощо).

В якості програмного засобу для проведення досліджень було використано інструмент для візуалізації даних, машинного навчання та інтелектуального аналізу даних Orange (рис. 1).

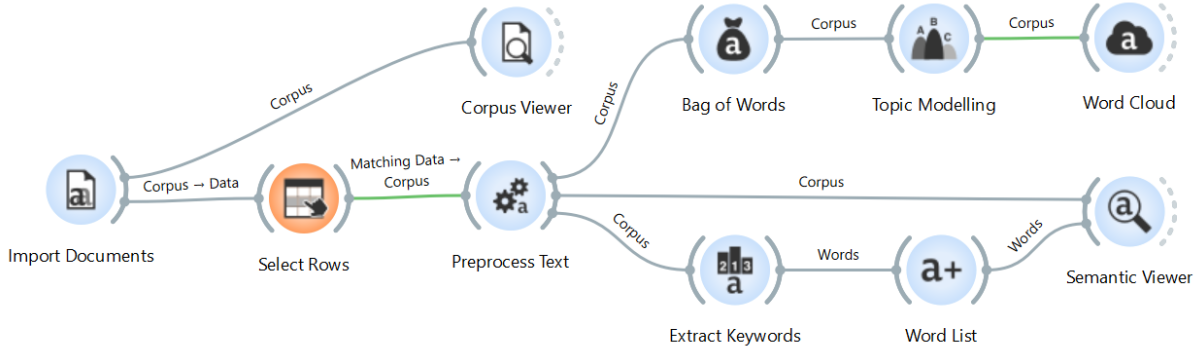

Рисунок 1 – Схема експерименту в Orange

Віджет «Preprocess Text» виконує попередню обробку тексту, а саме: заміна великих букв у всього тексті маленькими, видалення діакритичних знаків та наголосів в тексті, видалення html тегів та видалення url адрес; розбиття текстів на більш дрібні компоненти (слова, речення, біграми) без збереження знаків пунктуації; скорочення слів до основи, нехтуючи суфіксами чи закінченнями (стемінг), а також приведення схожих словоформ до їх основної словникової форми (лематизація); видалення непотрібних символів та слів без сенсу (наприклад, артиклі, прийменники, сполучники, частки та ін.). Віджет «Bag Of Words» формує набір документів, кожен з яких містить тільки найбільш уживані, характерні для документу слова та словосполучення. Віджет «Topic Modeling» знаходить найбільш характерні для всіх документів набори слів. Віджет «Extract Keywords» дозволяє виділити найбільш уживані (ключові) слова для набору документів. Віджет «Word List» призначений для створення та об'єднання списків слів для семантичного аналізу. Кількісна оцінка співпадіння тематик набору текстів та набору ключових слів обчислюється на рівні кожного речення, як максимальна косинусна подібність між SBERT поданням речення та ключовими словами.

Результати класифікації електронних листів дослідних користувачів за набором ключових слів, що мають ознаки інсайдерської діяльності, наведено в табл. 1.

Таблиця 1 – Результати пошуку інсайдерів серед 10 користувачів датасету «The Enron Email Dataset»

Тег «POI-true»	Співпадіння з набором ключових слів, %
LAY KENNETH	31.6
SKILLING JEFFREY	25.7
FASTOW ANDREW	46.7
DELAINEY DAVID	35.7
BOWEN RAYMOND	33.1
Тег «POI-false»	Співпадіння з набором ключових слів, %
ALLEN PHILLIP	16.5
BLACHMAN JEREMY	29.1
HAYES ROBERT	21.7
LOCKHART EUGENE	30.5
WALTERS GARETH	18.9

Як можна бачити, заявленої точності в 90 % ми не досягли, але можна зробити висновок про те, що застосування семантичного аналізу до текстів електронних листів має перспективи. Адже 3-х з 5-ти інсайдерів можна вважати було виявлено (якщо за граничну межу прийняти 32 %). Так само для 3-х з 5-ти звичайних співробітників значення ступеня співпадіння набагато менше граничної межі. Таким чином, можна говорити про точність детектування інсайдерських дій у 60 %. Правда слід зазначити, що експерименти проведені на досить невеликому наборі даних

Список використаних джерел:

1. Cappelli, D., Moore, A. & Trzeciak, R., 2012. The CERT Guide to Insider Threats. Westford, Massachussetts

2. Chi, H., Prodanoff, Z. G., Scarllet, C. & Hubbard, D., 2016. Determining Predisposition to Insider Threat Activities by using Text Analysis. Future Technologies Conference, pp. 985-990.

3. Young, W. T., Memory, A., Goldberg, H. G. & Senator, T. E., 2014. Detecting Unknown Insider Threat Scenarios. 2014 IEEE Security and Privacy Workshops, pp. 277-288.

4. A machine learning approach to detect insider threats in emails caused by human behavior. Dissertation by Antonia Michael. University Of Pretoria Pretoria, South Africa. 2020.

5. The Enron Email Dataset. URL: https://www.cs.cmu.edu/~enron/ (дата звернення: 01.12.2024)