УДК 004.056.5

СЕНТИМЕНТАЛЬНИЙ АНАЛІЗ ЯК ІНСТРУМЕНТ ДЛЯ ВИЯВЛЕННЯ ШКІДЛИВОГО КОНТЕНТУ

Шедін Д.А.
e-mail: dmytro.shedin@nure.ua
Харківський національний університет радіоелектроніки,
каф. ІКІ ім. В.В. Поповського
м. Харків, Україна

This work is devoted to sentiment analysis methods to detect malicious content in personal communications like social networks, e-mail and instant messengers. The features of manipulative techniques in phishing messages, fake pranks and spam messages are considered. A model is proposed that combines the analysis of emotional coloring with the identification of linguistic patterns characteristic of malicious content of a personal nature. Also, it shows where the research results can be applied.

В епоху цифрових комунікацій користувачі щодня стикаються з різноманітними видами шкідливого контенту: фішинговими атаками, фейковими розіграшами, спам-повідомленнями та іншими формами маніпуляцій. Сучасні системи фільтрації шкідливого контенту переважно базуються на аналізі ключових слів, структурних особливостей повідомлень та технічних параметрів, як-от IP-адреси чи домени. Проте ефективність таких підходів є обмеженою, особливо коли йдеться про цільові атаки та новітні методи обходу захисних механізмів, що постійно еволюціонують.

Сентиментальний аналіз же є додатковим інструментом, який може значно підвищити точність розпізнавання шкідливого контенту. Він традиційно визначається як метод виявлення суб'єктивних характеристик тексту – емоційного забарвлення, оцінок, установок та намірів автора [1]. У контексті виявлення зловмисної інформації в особистих комунікаціях особливий інтерес становлять:

- емоційні тригери: шкідливий контент часто апелює до базових емоцій (страх, жадібність, цікавість, терміновість тощо), щоб спонукати користувача до необдуманих дій;
- розбіжність тональності: невідповідність емоційного забарвлення текстовому контексту або заявленій меті повідомлення (наприклад, надмірна позитивність у повідомленнях про виграші або загрозлива тональність у повідомленнях від «служби підтримки»);
- маніпулятивні наративи: використання специфічних сценаріїв та наративних структур для введення користувача в оману.

Гіпотеза полягає в тому, що шкідливий контент у персональних комунікаціях має характерні патерни емоційного забарвлення, які можуть бути виявлені методами сентиментального аналізу та використані для підвищення точності ідентифікації такого контенту.

Для ефективного визначення необхідно враховувати специфіку різних типів шкідливого контенту.

Фішингові повідомлення характеризуються створенням атмосфери терміновості, імітацією офіційного стилю з емоційними тригерами, комбінаціями погроз та обіцянок розв'язати проблему. Прикладом такого є: «ТЕРМІНОВО! Ваш банківський рахунок заблоковано через підозрілу активність! Для відновлення доступу необхідно негайно підтвердити вашу особу, перейшовши за посиланням: [шкідливе посилання]. Невиконання цієї вимоги призведе до повного блокування та можливої втрати коштів. Наша служба безпеки готова допомогти вам вирішити цю проблему – просто введіть ваші дані на сайті».

Фейкові розіграші мають надмірно позитивне емоційне забарвлення, апеляції до почуття удачі та ексклюзивності, створення відчуття обмеженості пропозиції в часі. Одним із прикладом такого повідомлення є: «ВАУ! ВІТАЄМО! Ви стали 1000-м відвідувачем нашого сайту і виграли НОВИЙ ТЕЛЕФОН Pro Max! Це ЕКСКЛЮЗИВНА пропозиція тільки для вас! Поспішайте отримати свій приз протягом наступних 30 хвилин – така удача випадає раз у житті! Просто заповніть форму за посиланням і ваш неймовірний подарунок буде відправлено вам!».

Спам-повідомлення мають гіперболізоване представлення переваг продукту чи послуги, використання мовних конструкцій, що створюють відчуття невідкладності, надмірне застосування окличних речень та емоційно забарвленої лексики. Повідомленням такого характеру може бути: «РЕВОЛЮЦІЙНИЙ АНТИВІРУС! ЗАХИСТІТЬ свій комп'ютер від УСІХ атак СЬОГОДНІ! Наш супер-захист блокує 100% ВСІХ загроз, які пропускають звичайні антивіруси! МІЛЬЙОНИ користувачів вже ВРЯТОВАНІ від кібератак! ЛИШЕ ЗАРАЗ знижка 75%! Встановіть протягом 24 ГОДИН і отримайте БЕЗКОШТОВНИЙ VPN на рік! НЕ РИЗИКУЙТЕ своїми даними! НЕГАЙНО завантажте наш НАДІЙНИЙ захист!!!».

Шахрайські схеми в соціальних мережах містять маніпуляції довірою через фальшиву емпатію, поєднання позитивних емоційних тригерів з нав'язливими закликами до дії, створення штучного емоційного зв'язку з потенційною жертвою. Прикладом такої є: «Ексклюзивна інформація! Тільки для обраних! Криптовалюта XYZ зросте на 500% за тиждень! Я вже заробив 50 000$ і хочу поділитися секретом з тобою. Приєднуйся до нашої приватної групи зараз! Місць обмежено! Перші 10 учасників отримають безкоштовну консультацію. Пиши в приватні!».

Для досягнення поставленої мети щодо визначення таких повідомлень використовується багаторівневий підхід, що включає:

1) Побудову спеціалізованого корпусу даних: збір та анотування вибірки шкідливих повідомлень з різних джерел (електронна пошта,

соціальні мережі, месенджери тощо), класифікація за типами шкідливого контенту, розмітка емоційних тригерів та маніпулятивних елементів.

2) Багаторівневий сентиментальний аналіз: визначення загальної тональності повідомлення, виявлення емоційних тригерів на рівні окремих фраз, аналіз дисонансу між заголовком та основним текстом, оцінка невідповідності тональності контексту.

3) Інтеграцію з традиційними методами виявлення: розробка гібридної моделі, що поєднує сентиментальний аналіз із лексичними та структурними підходами, одним з яких є власна розроблена система «Adressant» [2].

4) Застосування методів машинного навчання для класифікації повідомлень.

5) Тестування в умовах реального використання.

Результати дослідження, а також сама система можуть бути використані для розробки комплексних рішень з інформаційної безпеки, удосконалення наявних антиспам-фільтрів, антивірусів, а також для освітніх програм з цифрової грамотності та захисту від інформаційних маніпуляцій.

Список використаних джерел:

1. Medhat W., Hassan A., Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal. URL: https://www.sciencedirect.com/science/article/pii/S2090447914000550. (дата звернення 01.03.2025).

2. Шедін Д.А. Програмне забезпечення для цифрового криміналістичного аналізу повідомлень електронної пошти. Радіоелектроніка та молодь у XXI столітті: каталог виставки технічної творчості молоді, м. Харків, 10–12 трав. 2023 р. / Харків. нац. ун-т радіоелектроніки.