РОЗРОБКА СИСТЕМИ ФІЛЬТРАЦІЇ ЕЛЕКТРОННОЇ ПОШТИ НА ОСНОВІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Даншин В.В., Балагура Д.С.
Харківський національний університет радіоелектроніки, Харків, Україна

Сучасне інформаційне суспільство характеризується стрімким розвитком цифрових технологій і зростанням обсягів електронного листування. Електронна пошта є одним із найпоширеніших засобів комунікації в бізнесі, державному секторі та повсякденному житті [1]. Разом із цим різко зростає кількість кіберзагроз, пов'язаних із поштовими повідомленнями - спам, фішингові атаки, шкідливі вкладення, соціальна інженерія тощо [2]. Традиційні системи фільтрації електронної пошти, що базуються на фіксованих правилах або простих статистичних моделях, уже не здатні ефективно протидіяти сучасним атакам, які постійно змінюють структуру та зміст повідомлень. У цих умовах актуальним стає використання технологій штучного інтелекту (ШІ), зокрема великих мовних моделей (LLM) [3], які можуть здійснювати семантичний аналіз, розуміти контекст і логіку тексту, а також самонавчатися на основі нових прикладів.

Метою роботи є розроблення інтелектуальної системи фільтрації електронної пошти із застосуванням сучасних моделей машинного навчання та архітектури Retrieval-Augmented Generation [4]. У роботі розглядаються актуальні загрози електронній пошті, такі як спам, фішинг, spear-phishing, шкідливі вкладення та автоматизовані ШІ-згенеровані повідомлення. Аналізуються класичні методи фільтрації black/white списки, rule-based, баєсів фільтр та порівнюються з підходами на основі трансформер-моделей і векторного пошуку. Реалізується прототип системи, що поєднує нейронні мережі, динамічне формування правил і гібридну обробку даних з елементами RAG-архітектури. Результатом дослідження є створення адаптивної системи, здатної виявляти спам і фішингові повідомлення з високою точністю, пояснювати свої рішення та динамічно оновлювати правила класифікації. Створення UI дизайну для користувача та адміністратора для коригування роботи системи і аналізу роботи ШІ. Отримані результати підтверджують доцільність впровадження ШІ у сферу кіберзахисту електронної пошти.

Список літератури

1. Cloudflare. Email Security Threat Report, 2024. [Електронний ресурс]. – Режим доступу: https://blog.cloudflare.com/radar-2024-year-in-review/
2. Голубничий, Д.Ю., Сєвєрінов, О.В., Коломійцев, О.В., та інш. (2021). Аналіз сучасних загроз в інформаційних системах за складовими загроз: кібербезпеки, інформаційної безпеки та безпеки інформації.
3. Large language models [Електронний ресурс]. – Режим доступу: https://link.springer.com/article/10.1007/s10462-024-10888-y
4. Retrieval-Augmented Generation [Електронний ресурс]. – Режим доступу: https://www.sciencedirect.com/science/article/pii/S1877050924021860