

к.філол.н., доц. Бабич Б., (Університет Лідсу, Великобританія)
phd, доц. Атвелл Е., (Університет Лідсу, Великобританія)
b.babych@leeds.ac.uk; e.s.atwell@leeds.ac.uk

ЛІНГВІСТИЧНЕ ЗАБЕЗПЕЧЕННЯ СИСТЕМИ ВИЛУЧЕННЯ ІНФОРМАЦІЇ ДЛЯ ЗАВДАНЬ АВТОМАТИЧНОГО ВІДСЛІДКОВУВАННЯ ПРОПАГАНДИ ТЕРОРИЗМУ

У доповіді представлено систему, що розробляється для завдань багатомовного пошуку, вилучення та аналізу інформації про терористичні загрози. Система працює в реальному часі, відслідковуючи статті, повідомлення, коментарі та блоги в мережі Інтернет, що містять пропаганду тероризму, заклики до насильства та нетерпимості. Наш підхід поєднує технології машинного перекладу (МП), вилучення інформації (ВІ) та обчислення схожості текстів, що містять ознаки пропаганди, у корпусі електронних ЗМІ. Лінгвістичне забезпечення системи розробляється дослідницькими групами штучного інтелекту та перекладознавства Університету м. Лідс і є частиною проекту поданого на грант програми Horizon2020 спільно з групою поліцейських управлінь ряду міст та провідними компаніями в галузі безпеки з 6 країн ЄС. Проект поєднує технології інших учасників, такі як автоматичний аналіз рукописного тексту, фото та відеоматеріалів, виявлення модифікованих фотографій, аналіз графів зв'язків і повідомлень у соціальних мережах.

Технологія ВІ в галузі безпеки традиційно використовувалася з метою виявлення незаконної терористичної діяльності, а саме реальних загроз нападів; але з іншого боку, пропаганда тероризму та релігійного фундаменталізму, хоча і не є незаконними у демократичних країнах, дедалі більше стають потужними засобами ідеологічного виправдання терористичної діяльності, радикалізації та набору найманців, створення локальних терористичних осередків, діяльність яких набагато важче виявити. Новим елементом у пропонованому нами підході є відслідковування терористичної пропаганди та дезінформації в електронних мережах у реальному часі, що дозволяє аналітикам виявляти джерела поширюваних повідомлень, оцінювати рівень їхньої загрози, знаходити правдиві факти, повідомлення і ресурси для протидії інформаційним нападам.

Програмно-лінгвістичне забезпечення включає систему статистичного МП, створену на корпусі перекладів з арабської, української та російської мови на англійську, та систему ВІ для англійської мови, що виявляє шаблони пропагандистських повідомлень і обчислює їхню семантичну схожість у різних джерелах. Шаблони системи ВІ визначено у термінах конфігурацій лексем та частиномовних кодів, генерованих системою автоматичного морфологічного аналізу, а також елементів формальної моделі предметної галузі (онтології), що містить ідеологічні та релігійні концепти, які можуть використовуватися терористичними групами. Зараз напрямками нашого дослідження є автоматизована побудова онтологій з корпусу текстів, автоматичний логічний вивід в онтологіях з використанням дескриптивної логіки, визначення схожості шаблонів та описів подій при багатотекстовому ВІ, моделювання діалектної та графологічної варіативності, скорочень та помилок, метафоричної мови та евфемізмів у джелелах і повідомленнях, що містять пропаганду тероризму.