

РЕФЕРАТ

Магістерська дисертація: 94 с., 22 рис., 19 табл., 1 додаток, 20 джерел.

Актуальність. Пропаганда існує в традиційних засобах масової інформації протягом багатьох років, і з розвитком Інтернету поступово проникає і у соціальні медіа. Це обумовлено щоденним збільшенням текстової інформації на просторах всесвітньої мережі інтернет. Дійсно, пропаганда настільки потужна, що всі схильні до цього. Основним каналом, по якому соціум приймає пропаганду, є засоби масової комунікації. Статистика переконливості засобів масової інформації вражаюча. Поширення інформації - це індустрія з оборотами більш ніж в 400 мільярдів доларів на рік, з яких 206 мільярдів витрачаються на масову пропаганду. Кожна людина мислить вербально і тому в тій чи іншій мірі піддається впливу, чим вміло користуються фахівці з продажу, політики, медійники, шахраї, організатори сект, спецслужби і терористи.

Отже, необхідною задачею є створення інструменту інформаційно-психологічної безпеки, призначеного для перевірки текстового контенту на наявність спеціальних лінгвістичних конструкцій та оборотів, які сприяють некритичному аналізу інформації.

Зв'язок роботи з науковими програмами, планами, темами. Робота виконувалась на кафедрі автоматизованих систем обробки інформації та управління Національного технічного університету України «Київський політехнічний інститут ім. Ігоря Сікорського» в рамках теми «Методи та технології високопродуктивних обчислень та обробки надвеликих масивів даних». Державний реєстраційний номер 0117U000924.

Метою дослідження є створення алгоритму, призначеного для перевірки текстового контенту на наявність спеціальних лінгвістичних конструкцій та оборотів, які сприяють некритичному аналізу інформації.

Для досягнення поставленої мети необхідно виконати наступні завдання:

- проаналізувати алгоритми та методи машинного навчання для вирішення задачі автоматичної класифікації текстів;

- обрати модель представлення текстової інформації в класифікаторі;
- розробити алгоритм попередньої обробки текстів відповідно до обраної моделі представлення текстової інформації;
- розробити модифікований метод класифікації текстової інформації;
- виконати програмну реалізацію розробленого алгоритму автоматичної класифікації текстової інформації;
- підготувати дані для оцінки якості класифікації;
- провести дослідження ефективності розробленої інформаційної технології.

Об'єктом дослідження є процес класифікації текстових даних.

Предметом дослідження є методи класифікації текстових даних для виявлення пропаганди.

Методами дослідження є методи машинного навчання, які базуються на методах text mining.

Наукова новизна отриманих результатів. Розроблено модифікований метод класифікації текстових даних для виявлення текстових даних, що використовує попередній сентиментальний аналіз, латентне розміщення Дирихле та алгоритм TextRank. Алгоритм TextRank було покращено завдяки врахування позиції слова у тексті.

Публікації. Результати досліджень опубліковані в журналі «Сучасні напрями розвитку інформаційно-комунікаційних технологій та засобів управління» [1], опубліковано в тезах наукової конференції студентів, магістрантів та аспірантів «Інформатика та обчислювальна техніка» – ІОТ-2018 [2], опубліковані в журналі “World Congress “Aviation in 21st century”[3].

МАШИННЕ НАВЧАННЯ, ТЕХТ MINING, ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ, ТЕКСТОВИЙ КЛАСИФІКАТОР, АНАЛІЗ КОНТЕНТУ, АЛГОРИТМИ КЛАСИФІКАЦІЇ