АНОТАЦІЯ

Стецик О.А. Інтелектуальна високонавантажена розподілена система обробки даних в соціальних мережах. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 126 «Інформаційні системи та технології». – Київський національний університет будівництва і архітектури. – Київ, 2024.

Наукова новизна одержаних результатів. Наукова новизна дисертаційної роботи полягає у тому, що:

вперше розроблено:

- модель НМКТуСН, що є нейронною мережею для класифікації тексту у стрічці новин на основі модифікації гібриду згорткової нейронної мережі та двосторонньої мережі довгої короткочасної пам’яті, яка ефективно поєднує архітектурні рішення, оптимізовані гіперпараметри та натреновані наперед векторні вбудування слів за допомогою нейромереж з трансформер архітектурою, для того, щоб класифікувати текст ефективно та швидко;

- метод зменшення ефекту «бульбашки» через диверсифікацію стрічки новин, який грунтується на класифікації розробленою нейромережею новин за конкретними дискусійними категоріями, і подальшій кластеризації текстів в цих дискусійних категоріях; запропонований метод диверсифікації стрічки новин, сприяє більш тонкому ознайомленню з різними точками зору, щодо дискусійної категорії;

удосконалено:

- метод виявлення спаму і пропаганди, за рахунок тренування НМКТуСН, в напрямку отримання балансу між точністю і швидкістю класифікації;

- архітектуру системи соціальних мереж в напрямку наближення швидкості класифікацій текстів і генерації диверсифікованої стрічки новин до режиму реального часу, шляхом комбінації масштабованих технологій і натренованої НМКТуСН;

набули подальшого розвитку:

- концепція створення гібридних систем, що заснована на моделях штучного інтелекту та машинного навчання в напрямку обробки даних для отримання чистішого і різноманітнішого середовища в соціальних мережах.

Основний зміст дисертаційної роботи.

Дисертаційну роботу присвячену проблемам які виникають у високонавантажених розподілених системах соціальних мереж, таким як: спам, пропаганда, та ефект «бульбашки», та їх вирішенню у реальному часі та при великій кількості користувачів.

За результатами дослідження розроблено:

- детальну архітектуру високонавантаженої розподіленої системи соціальної мережі, яка в реальному часі маркує пости, як спам чи пропаганда, та вказує конкретну пропагандистську техніку, яка використовується;

- метод диверсифікації стрічки новин, який дозволяє зменшення ефекту «бульбашки».

У першому розділі «Аналіз сучасних інтелектуальних високонавантажених розподілених систем соціальних мереж» досліджено сучасні інтелектуальні системи та проблеми, що у них виникають; виконано аналіз існуючих моделей і методів боротьби із спамом та пропагандою; виконано аналіз причин виникнення ефекту «бульбашки» та моделей і методів боротьби з ним; проаналізовано архітектурні вразливі місця розподілених систем соціальних мереж; досліджено причини виникнення вразливостей, таких як: відмову серверів, раптовий наплив великої кількості користувачів, велику кількість одночасних запитів; проаналізовано методи виявлення та усунення несправностей у розподілених системах; обгрунтувано необхідність створення ефективних моделей і методів розпізнавання спаму і пропаганди та боротьби з ефектом «бульбашки».

У другому розділі «Моделі та методи виявлення спаму, пропаганди і зменшення ефекту «бульбашки» в соціальних мережах» описано переваги та недоліки застосування машинного навчання і нейромереж для класифікації спаму, пропаганди та зменшення ефекту «бульбашки»; запропоновано модель нейронна мережа для класифікації тексту у стрічці новин (НМКТуСН) на основі модифікації

гібриду згорткової нейронної мережі та мережі довгої короткочасної пам’яті, яка ефективно поєднує архітектурні рішення, оптимізовані гіперпараметри і натреновані наперед векторні вбудування слів на основі нейромереж з трансформер архітектурою, для ефективної та швидкої класифікації тексту; удосконалено метод класифікації спаму і пропаганди за допомогою використання моделі НМКТуСН; запропоновано метод боротьби з ефектом «бульбашки» на основі моделі НМКТуСН, та кластеризації текстів.

У третьому розділі “Створення моделі високонавантаженої розподіленої системи соціальної мережі” описано необхідні операції користувачів у соціальних мережах; удосконалено масштабовану модель для соціальної мережі; описано напрямки розширення даної моделі, при збільшенні кількості функціоналу; визначено вразливі місця запропонованої моделі; під час моделювання даної системи особлива увага приділяється швидкодії даної системи, та її пропускній здатності; інтегровано модель високонавантаженої системи соціальної мережі з методами класифікації постів на пропаганду та спам; інтегровано модель з методом диверсифікації стрічки новин для зменшення ефекту «бульбашки» в соціальних мережах; покращено систему в напрямку збільшення швидкодії за рахунок використання мережі доставки повідомлень та кешування, додано моніторинг швидкості запитів та наявність помилок.

У четвертому розділі: «Експериментальні дослідження отриманих моделей та методів**»** розроблено інструментальні засоби для класифікації текстів на спам і пропаганду, та кластеризації текстів; проведено порівняння результатів класифікації текстів на спам і пропаганду за допомогою різних моделей і методів машинного навчання та нейронних мереж на основі оцінок влучності, повноти, точності та f-міри; проведено експерименти для кластеризації текстів по дискусійних темах.

Ключові слова: штучний інтелект, машинне навчання, нейронні мережі, управління знаннями, великі дані, швидкодія, пропаганда, спам.

ABSTRACT

Stetsyk O.A. Intelligent high-load distributed data processing system in social networks.

Dissertation for the degree of Doctor of Philosophy in the speciality 126 "Information Systems and Technologies." - Kyiv National University of Construction and Architecture.

Scientific novelty of the results. The scientific novelty of the dissertation is that:

for the first time developed:

- a model NMfTCiNF, which is a neural network for text classification in a news feed based on a modification of a hybrid of a convolutional neural network and a bidirectional long short-term memory network, which effectively combines architectural solutions, optimised hyperparameters and pre-trained vector word embeddings using neural networks with transformer architecture in order to classify text efficiently and quickly;

- a method for reducing the "bubble" effect through news feed diversification, which is based on the classification of news by the developed neural network into specific discussion categories, and the subsequent clustering of texts in these discussion categories; the proposed method of diversifying the news feed contributes to a more subtle introduction to different points of view regarding a discussion category;

improved:

- the method of detecting spam and propaganda, by training NMfTCiNF, in the direction of obtaining a balance between accuracy and speed of classification;

- the architecture of the social networking system in the direction of bringing the speed of text classifications and generation of a diversified news feed closer to real-time, by combining scalable technologies and trained NMfTCiNF;

were further developed:

- the concept of creating hybrid systems based on artificial intelligence and machine learning models in the direction of data processing to obtain a cleaner and more diverse environment in social networks.

The main content of the thesis.

The dissertation is devoted to the problems that arise in highly loaded distributed social networking systems, such as spam, propaganda, and the bubble effect, and their solution in real time and with a large number of users.

Based on the results of the study, we have developed

- a detailed architecture of a highly loaded distributed social network system that labels posts as spam or propaganda in real time and indicates the specific propaganda technique used;

- a method of diversifying the news feed to reduce the bubble effect.

The first chapter, "Analysis of Modern Intelligent Highly Loaded Distributed Social Networking Systems", examines modern intelligent systems and the problems they face; analyses existing models and methods of combating spam and propaganda; analyses the causes of the ‘bubble’ effect and models and methods of combating it; analyses the architectural vulnerabilities of distributed social networking systems; investigates the causes of vulnerabilities, such as server failure, sudden influx of a large number of users, large number of simultaneous requests; analyses methods of detecting and eliminating faults in distributed systems; substantiates the need to create effective models and methods for recognising spam and propaganda and combating the "bubble" effect.

The second chapter, "Models and methods for detecting spam, propaganda and reducing the bubble effect in social networks," describes the advantages and disadvantages of using machine learning and neural networks to classify spam, propaganda, and reduce the bubble effect; a model of a neural network for text classification in a news feed (NMfTCiNF) based on a modification of a hybrid of a convolutional neural network and a long short-term memory network, which effectively combines architectural solutions, optimised hyperparameters and pre-trained vector word embeddings based on neural networks with a transformer architecture, for effective and fast text classification; improved the method of classifying spam and propaganda by using the NMfTCiNF model; proposed a method for combating the "bubble" effect based on the NMfTCiNF model and clustering texts.

In the third chapter, "Models and methods for detecting spam, propaganda and reducing the bubble effect in social networks", the necessary operations of users in social

networks are described; the scalable model for the social network is improved; the directions of expansion of this model are described, with an increase in the number of functions; the vulnerabilities of the proposed model are identified; when modelling this system, special attention is paid to the performance of this system and its throughput; the model of a highly loaded social network system is integrated with classification methods.

In the fourth chapter: "Experimental studies for the obtained models and methods" develops tools for classifying texts into spam and propaganda and clustering texts; compares the results of classifying texts into spam and propaganda using different models and methods of machine learning and neural networks based on the accuracy, completeness, precision and f-measure; conducts experiments for clustering texts on controversial topics.

Keywords: artificial intelligence, machine learning, neural networks, knowledge management, big data, performance, propaganda, spam.