УДК 004.91
А. Козак, С. Дячук, канд. техн. наук; доц.
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ОБРОБКА ПРИРОДНЬОЇ МОВИ ДЛЯ ВИЯВЛЕННЯ І ЗАПОБІГАННЯ МАСОВОЇ ДЕЗІНФОРМАЦІЇ

UDC 004.91
A. Kozak, S. Dyachuk, Ph.D.; Assoc. Prof.

NATURAL LANGUAGE PROCESSING FOR DETECTING AND PREVENTING MASS DISINFORMATION

Дезінформація – це тип інформації, який створюється і поширюється з наміром введення кінцевого користувача в оману стосовно реального стану справ. Постійне споживання дезінформації призводить до викривленої реальності, через це поширення неправдивих новин зазвичай відбувається у пропагандистських, військових або комерційних цілях.

На сьогодні яскраві приклади дезінформації щоденно зустрічаються в соціальних мережах. Цифрові гіганти Meta, Google, Twitter володіють платформами де щоденно поширюються мільярди новин, тому повинен існувати механізм що забезпечує їх достовірність.

Ядром інформаційної системи для виявлення і запобігання масової дезінформації є класифікатори що працюють з природною мовою (природна мова являє собою сукупність певних звуків та символів загальноприйнятих у певному суспільстві, за допомогою яких люди виражають свої думки). Станом на 2021 рік людством створено 50.5 екзабайт даних. Аналітики прогнозують, що до 2025 року загальна кількість даних зросте до 175 екзабайт. Постійне зростання кількості та складності даних спонукає розробляти інструменти та проводити дослідження у цій галузі, адже потенційно можна тримати значну кількість корисної інформації аналізуючи такі дані. Однак через високу складність аналізу та структурування таких даних їх зазвичай ніяк і ніде не використовують.

Ще з перших версій електронно-обчислювальних машин, програмісти намагались навчити комп’ютер розуміти природну мову. Причина досить зрозуміла – за тисячі років люди згенерували таку велику кількість інформації такого типу, що не звертати увагу просто не можливо. На жаль комп’ютери не можуть в певній мірі розуміти природну мову так, як це роблять люди, але їх можливості значно розширились за останні два десятки років, з розвитком обробки природної мови (Natural Language Processing, NLP). NLP – це одна із галузей штучного інтелекту, яка займається аналізом та синтезом природної мови. Останні розробки у сфері NLP доступні через відкриті бібліотеки spaCy та NLTK, на багатьох мовах програмування, зокрема Python.