УДК 004.8

Трохимчук О.В., Пасічник О.А., Поплавська О.А , Міхалевський В.Ц.

Хмельницький національний університет

ПІДХІД ДО ОЦІНЮВАННЯ ВІДПОВІДНОСТІ ХЕШТЕГІВ КОРОТКИМ ТЕКСТАМ ЗАСОБАМИ NLP

Розглянуто задачу автоматизованого оцінювання відповідності хештегів коротким текстам соціальних мереж в умовах змішаних мов, неформальних конструкцій та маніпулятивного маркування контенту. Запропоновано підхід, що базується на контекстних мовних моделях типу трансформерів і побудові спільного семантичного простору для текстів і хештегів, попередньо сегментованих і перетворених на фрази природної мови. Показано можливість інтеграції підходу в модулі модерації, рекомендаційні сервіси й аналітичні системи для виявлення нерелевантних і маніпулятивних тегів та підвищення якості тегування.

The paper addresses the task of automated assessment of hashtag relevance to short social media texts under conditions of mixed languages, informal style and manipulative tagging. The proposed approach relies on transformer-based contextual language models and a joint semantic space for texts and hashtags, which are first segmented and converted into natural-language phrases. The approach can be integrated into moderation modules, recommender systems and analytical platforms to detect irrelevant or manipulative hashtags and improve the overall quality of tagging.

Поширення соціальних мереж та мікроблогів сформувало окрему «мову» коротких повідомлень, у якій хештеги виконують роль компактних маркерів тематики, настрою, події та цільової аудиторії [1]. Від коректності їх використання залежать якість пошуку, релевантність рекомендацій, прозорість аналітики та можливості автоматизованого моніторингу інформаційних потоків. Натомість практика свідчить про масове зловживання хештегами: від випадкового добору «модних» тегів до цілеспрямованого маніпулятивного маркування контенту. Це актуалізує задачу автоматизованого оцінювання відповідності хештегів коротким текстам, яка природно розв'язується засобами обробки природної мови.

Проблема є нетривіальною вже на рівні формалізації. Короткі тексти соціальних мереж містять розмовні конструкції, емодзі, скорочення, змішані мови, орфографічні відхилення, а хештеги поєднують словосполучення, сленг та елементи брендованої лексики [2]. Семантика хештегу не завжди збігається із буквальним змістом його компонентів: теги можуть позначати кампанію, мем, локальний

інформаційний контекст. Тому прості евристики на кшталт підрахунку перетину лем чи збігу окремих ключових слів дають обмежені результати. Потрібна модель, яка працює в спільному семантичному просторі для тексту й хештегу, враховує контекст уживання та здатна відрізняти тематичну релевантність від поверхневої схожості.

Мета роботи полягає у розробленні підходу до оцінювання відповідності хештегів коротким текстам на основі сучасних моделей NLP із використанням контекстних подань, що дає змогу формувати кількісні показники релевантності, придатні для інтеграції в системи модерації, рекомендаційні сервіси та аналітичні панелі. Для досягнення мети пропонується узгоджена послідовність етапів: формування корпусу, лінгвістичний препроцесинг, побудова семантичних подань текстів і хештегів, навчання моделей оцінювання відповідності та валідація результатів на експертно розмічених даних.

На рівні даних базовою сутністю є пара «короткий текст - множина хештегів». Для дослідження формується корпус постів із соціальних платформ, де хештеги використовуються як основний механізм індексування. Для частини записів здійснюється експертне маркування, коли кожен хештег віднесено до одного з рівнів: релевантний змісту, частково релевантний (дуже широкий або вторинний щодо основної теми) та нерелевантний або маніпулятивний. Така градація важлива для навчання моделей, які мають відрізняти цілком доречні теги від тегів, що лише формально не суперечать тексту, але не відображають його змістового ядра. Додатково фіксуються метадані, зокрема час публікації, домінантна мова, тип акаунта, що дозволяє у подальшому враховувати платформні та часові зсуви.

Лінгвістичний препроцесинг коротких текстів передбачає очищення від надлишкової технічної інформації, нормалізацію скорочень та приведення токенів до лематизованої форми із збереженням емотивних маркерів і ключових емодзі, які впливають на інтерпретацію. Особливу увагу приділено обробці змішаних текстів із поєднанням української, англійської та транслітераційних форм. Хештеги проходять окрему фазу сегментації: розбиття композитних тегів на складники, усунення декоративних символів, перетворення на фразу природної мови.

Ключовим компонентом підходу є побудова спільного семантичного простору для текстів і хештегів. Для цього застосовуються контекстні мовні моделі типу трансформерів, які навчені на багатомовних або спеціалізованих соціально-медійних корпусах. Короткий текст пропускається через модель для отримання векторного подання, що агрегує інформацію на рівні висловлювання, а хештег, попередньо перетворений на фразу, обробляється тим самим механізмом. У

результаті і текст, і хештег репрезентовано у спільному латентному просторі, в якому семантично узгоджені об'єкти розташовані близько один до одного.

Оцінювання відповідності реалізується двома взаємодоповнювальними способами. Перший базується на безпосередньому вимірюванні подібності між векторами тексту та хештегу за допомогою косинусної метрики та подальшій калібровці порогових значень на валідаційній вибірці. Цей підхід є обчислювально економним і підходить для масової первинної фільтрації. Другий спосіб розглядає пару «текст-хештег» як єдину послідовність і подає її на вхід моделі типу cross-encoder, яка навчається класифікувати ступінь відповідності на основі повного контексту, включно з лексичними, синтаксичними та прагматичними ознаками. Така модель краще захоплює складні випадки, зокрема іронію, метафоричне вживання та тематичні зсуви.

Навчання моделей здійснюється у напівконтрольованій постановці. Експертно розмічена частина корпусу використовується для супервізованого навчання класифікатора ступеня релевантності, тоді як невеликі вбудовані у корпус блоки даних без явних анотацій залучаються через контрастивні схеми: правильні пари «текст–хештег» протиставляються штучно згенерованим негативним прикладам, утвореним шляхом перестановки тегів між текстами схожої тематики. Це підсилює здатність моделі розрізняти не лише довільні нерелевантні поєднання, а й тонкі відмінності між близькими за тематикою, але семантично відмінними тегами.

Для оцінювання якості запропонованого підходу використовуються стандартні метрики класифікації, а також ранжувальні показники. На рівні окремих пар аналізуються точність, повнота та F_1-міра для кожного рівня релевантності, а також зважені усереднення за класами. На рівні множини хештегів для одного тексту розглядаються метрики, що характеризують порядок, у якому система пропонує або підтверджує теги, зокрема середня позиція релевантних хештегів та нормалізована кумулятивна оцінка приросту. Порівняння з базовими підходами, що використовують лише статистичні співзустрічі слів або поверхневі подібності, демонструє переваги контекстних моделей у розв'язанні задачі, особливо для коротких, емоційно насичених і семантично неоднозначних текстів.

Практичне значення розробленого підходу полягає у його здатності працювати як модуль у складі ширших інформаційних систем. У середовищі модерації контенту він дає змогу автоматично виявляти маніпулятивні чи «шумові» хештеги, які не відбивають зміст поста, але використовуються для привернення уваги або паразитування на трендових темах. У рекомендаційних сервісах оцінка відповідності може застосовуватись для очищення вхідних сигналів, що формують профіль користувача, зменшуючи вплив некоректних тегів на моделі персоналізації.

У системах аналітики підхід забезпечує побудову більш надійних карт тематичних кластерів, де хештеги виступають не довільними маркерами, а семантично верифікованими індикаторами дискусійних полів.

Окремим напрямом розвитку є поєднання оцінки відповідності з виявленням специфічних типів ризикового контенту. Нерелевантні або надто загальні хештеги часто супроводжують маніпулятивні повідомлення, приховану рекламу чи спроби обійти прості фільтри модерації. Інтеграція семантичних оцінок релевантності з іншими модулями аналізу, наприклад класифікаторами токсичності, дезінформації чи цифрової втоми, відкриває можливості для побудови багатовимірних профілів постів, де невідповідність хештегів є одним із сигналів ризику.

Підсумовуючи, запропонований підхід до оцінювання відповідності хештегів коротким текстам базується на використанні контекстних мовних моделей, спільного семантичного простору для текстів і тегів, а також поєднання подібнісних та класифікаційних процедур. Він враховує лінгвістичні особливості коротких, неформальних повідомлень, підтримує багатомовність, а також допускає гнучке налаштування порогів прийняття рішень залежно від завдань системи-споживача. Отримані результати свідчать про перспективність такого підходу для побудови інструментів контролю якості тегування, підвищення прозорості інформаційних потоків та зменшення впливу маніпулятивних практик у цифрових середовищах.

Перелік посилань

1. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation / D. Ghosal et al. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China. Stroudsburg, PA, USA, 2019. URL: https://doi.org/10.18653/v1/d19-1015

2. Wang L., Zhang L. Hawkes processes for understanding heterogeneity in information propagation on Twitter. Frontiers in Physics. 2022. Vol. 10. URL: https://doi.org/10.3389/fphy.2022.1019380