УДК 004.4

Домбровський Н.С., Скрипник Т.К., Вознюк Л.О.

Хмельницький національний університет

МЕТОД ІДЕНТИФІКАЦІЇ ПОДІЙ В УКРАЇНОМОВНИХ ТЕКСТАХ ЗАСОБАМИ ОБРОБКИ ПРИРОДНОЇ МОВИ

В даному дослідженні проводиться оцінка, наскільки текст тестового документа відповідає введеним користувачем ключовим токенам з урахуванням оцінки важливості слів у тексті. Для визначення цієї важливості можуть використовуватися два підходи: дисперсійна оцінка та метод BM-25. Передбачається автоматизований процес формування синонімічних рядів до введених користувачем токенів з метою розширення можливостей методу.

In this study, we evaluate the extent to which the text of the test document corresponds to the key tokens entered by the user, taking into account the importance of words in the text. To determine this importance, two approaches can be used: variance estimation and the VM-25 method. An automated process of forming synonymous series to the user-entered tokens is envisaged to expand the capabilities of the method.

Обробка природної мови (Natural Language Processing, NLP) – це міждисциплінарна галузь інформатики та мовознавства, що вивчає та розробляє методи та технології для взаємодії між комп'ютерами та людьми через природну мову. Вона включає в себе комплексний аналіз та обробку тексту та мови з метою автоматизації розуміння та генерації текстової інформації [1].

Метод ідентифікації подій в україномовних текстах засобами обробки природної мови є актуальною задачею в галузі обчислювальної лінгвістики та штучного інтелекту. Цей метод спрямований на автоматичне визначення подій, які відбуваються в тексті, зокрема, діяльності, подій, процесів та їхніх атрибутів. Ідентифікація подій є важливою для багатьох застосувань, включаючи аналіз новин, моніторинг соціальних мереж, розробку систем автоматичного розуміння тексту, аналізу семантики та багато інших областей [2].

Для специфічних задач NLP іноді використовуються поєднання декількох методів та моделей, а також аналіз контексту та зв'язків між словами та фразами. Застосування цих математичних методів дозволяє здійснювати складний аналіз текстів і виявляти події та інформацію, пов'язану з ними, що може бути корисним в різних додатках, включаючи аналітику соціальних медіа, пошукові системи, моніторинг новин тощо [3].

Для поставленого завдання було обрано метод дисперсійної оцінки, це статистичний метод для визначення різноманітності та варіабельності даних у

вибірці. Дисперсія грає важливу роль у багатьох аналітичних задачах, включаючи оцінку важливості слів у текстових документах для подальшого використання в методах обробки природної мови, таких як ідентифікація подій.

Дисперсійна оцінка базується на обчисленні середнього квадратичного відхилення (стандартного відхилення) даних від їхнього середнього значення. Вона вимірює, наскільки дані розподілені відносно середнього значення. Вища дисперсія вказує на більшу розкиданість даних, тоді як нижча дисперсія вказує на менший розкид [4].

Формула для обчислення дисперсії наведено у формулі:

$$\sigma^2 = \frac{\sum(xi - \mu)^2}{N}$$

де σ^2 – дисперсія, xi – кожне окреме значення вибірки, μ – середнє значення (середнє арифметичне) вибірки, N – кількість значень вибірки.

Основна ідея полягає в тому, що дисперсія вимірює, наскільки середнє значення відхиляється від кожного окремого значення вибірки. Велика дисперсія вказує на те, що дані мають великий розкид, тоді як мала дисперсія означає, що значення вибірки майже однакові.

Дисперсійна оцінка може бути важливим інструментом в контексті ідентифікації подій в текстах. Вона може використовуватися для визначення важливості слів у текстових документах, які можуть вказувати на ключові терміни або інформацію, що вказує на певні події чи теми [5]. Розрахунок дисперсійної оцінки може допомогти виокремити найбільш значущі слова, які слід аналізувати подальше в контексті ідентифікації подій.

При розробці методу ідентифікації подій в україномовних текстах за допомогою обробки природної мови, одним з важливих виборів є вибір підходу для оцінки важливості слів та токенів у текстах. У контексті поставленої задачі, дисперсійна оцінка виявляється більш вдалою стратегією. Дисперсійна оцінка більше ураховує різноманітність слів та токенів у тексті. Це означає, що вона надає важливість словам, які вносять різноманітність та багатогранність в текст, що є важливим аспектом ідентифікації подій. Таким чином, дисперсійна оцінка відображає різноманітність тексту, що може допомогти виокремити ключові слова та фрази, пов'язані з подіями.

Для реалізації методу ідентифікації подій в україномовних текстах необхідно чітко окреслити послідовність його виконання. Схема роботи методу наведена на рисунку 1.

Вхідні дані. На цьому кроці вхідні дані включають текстовий документ, який піддається обробці, і список токенів, які вводить користувач. Текстовий документ представляє собою послідовність слів або фраз, які містять інформацію про події.

Токенізація тексту. Токенізація – це процес розбиття тексту на окремі токени (слова, фрази, символи тощо). Цей процес виконується для підготовки

тексту до подальшого аналізу. Токени представлені як послідовність символів, і кожен токен має свій внутрішній ідентифікатор.

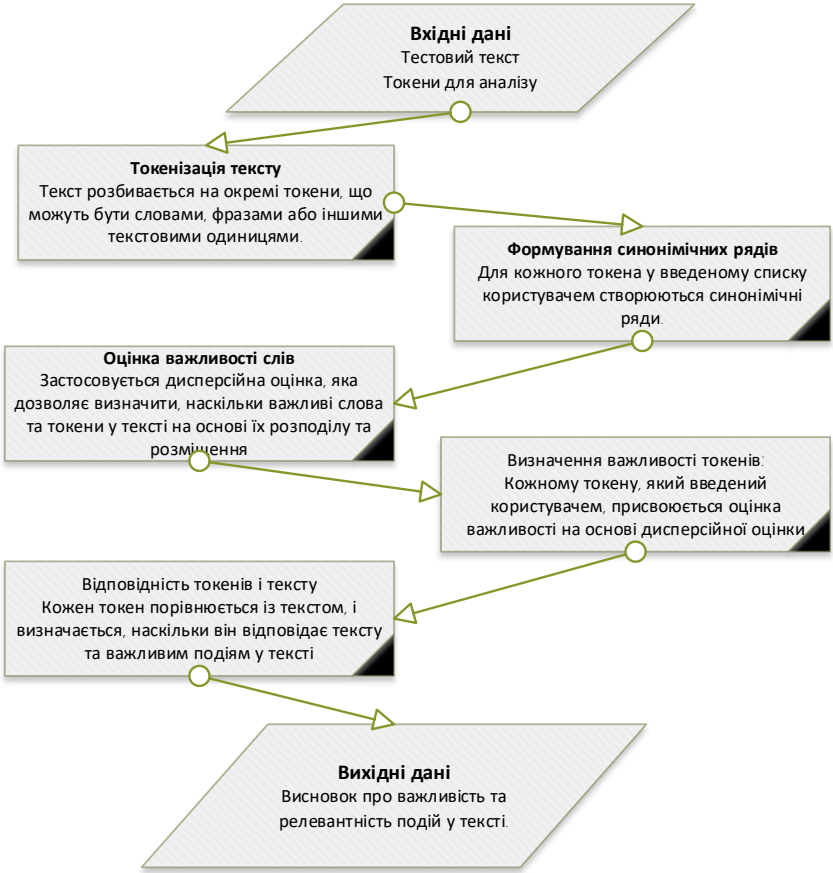

Рисунок 1 – Схема методу ідентифікації подій в україномовних текстах засобами обробки природної мови

Формування синонімічних рядів. Після токенізації для кожного токена, введеного користувачем, формуються синонімічні ряди. Синонімічний ряд - це набір інших слів або фраз, які мають схожий семантичний зміст або пов'язані за контекстом. Формування синонімічних рядів допомагає розширити можливості аналізу тексту.

Оцінка важливості слів. Для оцінки важливості слів у тексті використовується дисперсійна оцінка. Дисперсійна оцінка визначає, наскільки

слова розподілені у тексті та як вони пов'язані між собою. Ця оцінка допомагає визначити, які слова є ключовими для подій у тексті.

Визначення важливості токенів. Кожному токену, введеному користувачем, присвоюється оцінка важливості на основі дисперсійної оцінки. Оцінка важливості враховує розташування токену в тексті та його семантичний зв'язок з іншими словами. Токени, які мають високий рівень важливості, вважаються ключовими для подій у тексті.

Відповідність токенів і тексту. Кожен токен порівнюється із текстом з використанням отриманих оцінок важливості та дисперсійної оцінки. Це допомагає визначити, наскільки токен відповідає тексту і наскільки важливий для подій, які відображені в тексті.

Результати та висновок. На основі оцінок важливості та відповідності токенів тексту введеним користувачем токенам генерується висновок про важливість та релевантність подій у тексті. Ця інформація допомагає ідентифікувати та аналізувати події та їх контекст у тексті.

Отже, метод ідентифікації подій в україномовних текстах засобами обробки природної мови полягає в складному аналізі тексту та оцінці важливості токенів на основі дисперсійної оцінки. Цей метод дозволяє точно ідентифікувати та аналізувати події в тексті.

Таким чином, було проведено дослідження в галузі аналізу тексту, зокрема ідентифікації подій в україномовних текстах, запропоновано структуру методу, що може бути втілено в програмних реалізаціях.

Перелік посилань

1. An improved text mining approach to extract safety risk factors from construction accident reports
2. AL-NASSERI, Alya; ALI, Faek Menla; TUCKER, Allan. Investor sentiment and the dispersion of stock returns: Evidence based on the social network of investors. International Review of Financial Analysis, 2021, 78: 101910.
3. Text preprocessing for text mining in organizational research: Review and recommendations HICKMAN, Louis, et al. Text preprocessing for text mining in organizational research: Review and recommendations. Organizational Research Methods, 2022, 25.1: 114-146.
4. KONONOVA, Olga, et al. Opportunities and challenges of text mining in materials research. Iscience, 2021, 24.3. NA, X. U., et al. An improved text mining approach to extract safety risk factors from construction accident reports. Safety science, 2021, 138: 105216.
5. GRIES, Stefan Th. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. Research in Corpus Linguistics, 2021, 9.2: 1-33.