УДК 519.683, 519.767

АНАЛІЗ МЕТОДІВ ОПРАЦЮВАННЯ ЛІНГВІСТИЧНИХ СТРУКТУР У СИСТЕМАХ АВТОМАТИЗОВАНОЇ ОБРОБКИ ПРИРОДНОМОВНИХ ТЕКСТІВ

А.І. Воржевітіна[1], *Н.В. Шаронова*[2]

[1] *аспірант кафедри інтелектуальних комп'ютерних систем, НТУ «ХПІ», Харків, Україна*

[2] *професор кафедри інтелектуальних комп'ютерних систем, д-р. техн. наук, НТУ «ХПІ», Харків, Україна*

Anzhelika.Vorzhevitina@sgt.khpi.edu.ua

Обробка текстів природної мови – це одна з найважливіших областей досліджень у комп'ютерних науках (КН) і штучному інтелекті (ШІ). Обробка, як правило, передбачає перетворення текстових даних в числові, за допомогою яких комп'ютер отримує інформацію з природномовних текстів. З цієї мети розробляються технології NLP (Natural Language Processing), які включають методи опрацювання лінгвістичних структур, а саме Text Summarisation, Tokenisation, Parsing, Sentiment Analysis, Lemmatisation and Stemming, Stopwords Removal, TF-IDF (Term Frequency-Inverse Document Frequency), Named Entity Recognition (NER), Keyword Extraction, Word Embeddings, Topic Modelling. Аналіз методів опрацювання лінгвістичних структур у системах автоматизованої обробки природномовних текстів залишається актуальним завданням і має стратегічне значення, через необхідність удосконалення і розвитку для кращого розуміння та обробки природної мови [1].

Мета і завдання цієї роботи полягає в проведенні аналізу методів обробки лінгвістичних структур у системах автоматизованої обробки природномовних текстів для виявлення основних переваг і недоліків, отримання узагальненої інформації, яка надає змогу виявити найперспективніші методи оброблення лінгвістичних структур.

Головні переваги Text Summarisation – властивість генерації стислого, зв'язного підсумку тексту, через вилучення найважливішої інформації та ключових ідей з оригіналу. Недоліки – ризик втрати контексту та загальної змістовності, що може призвести до неповноцінної репрезентації оригіналу; може виявлятися менш ефективним при роботі зі складно структурованими текстами. Tokenisation розбиває текст на окремі одиниці (токени), які можуть бути словами, фразами або символами. Переваги – можливість підраховувати частотність слів, виявляти структуру речень і підготовувати вхідні дані для подальшого аналізу. Недоліки – обробка слів, що мають декілька значень, може призвести до неправильного поділу тексту; некоректна токенізація мов зі складною структурою. Parsing аналізує граматичну структуру речення для визначення синтаксичних зв'язків між словами. Переваги – ідентифікація граматичних структур мов, що дає можливість виявлення залежностей між словами та підвищує якість аналізу тексту. Недоліки – багатофункціональність та неоднозначність природних мов призводить до різних інтерпретацій, що ускладнює процес аналізу.

Sentiment Analysis визначає емоційний тон тексту – позитивний, негативний або нейтральний. Переваги – можливість швидкого та автоматизованого виявлення настроїв, що допомагає у розумінні відгуків, коментарів та загальної атмосфери тексту. Недоліки – обмежена точність у виявленні контекстуальних забарвлень і суб'єктивних висловлювань; різноманітність мовленнєвих засобів та культурні відмінності можуть ускладнювати процес правильної класифікації емоцій у тексті. Lemmatisation трансформує слова до словникової форми (lemma). Stemming зводить слова до їхньої

базової форми через видалення префіксів або суфіксів. Переваги – допомагає зменшити кількість варіацій слів і покращити зв'язність тексту та точність аналізу. Недоліки – потреба у найскладніших лінгвістичних моделях для правильного скорочення слів.

Stopwords Removal – метод видалення слів (стоп-слів), які не несуть суттєвого сенсу. Переваги – зменшення шуму і підвищення ефективності подальших аналізів. Недоліки – втрата цінної інформації при обробці, через видалення стоп-слів, які мають значення в контексті певного аналізу або важливу семантику, яку треба враховувати. TF-IDF дозволяє кількісно оцінити значимість терміна в документі або підбірці документів. Переваги – враховує частоту терміна в документі (частота терміна), і його рідкість у всій підбірці документів (зворотна частота документа). Недоліки – не враховує семантичні зв'язки між словами, що призводить до втрати контексту. Named Entity Recognition визначає та класифікує іменовані сутності в тексті. Переваги – вилучає структуровану інформацію з неструктурованого тексту. Недоліки – проблеми з взаємодією зі своєрідними контекстами й структуризацією мов.

Keyword Extraction автоматично вилучає важливі ключові слова чи фрази з тексту. Переваги – полегшення пошуку та навігації через великі обсяги даних, автоматична індексація документів. Недоліки – недостатнє розуміння контексту може призвести до вилучення непридатних ключових слів. Word Embeddings представляє слова у вигляді векторів у високовимірному просторі, фіксуючи їхні семантичні зв'язки. Переваги – здатність представляти семантику слів та контексту якісно та кількісно. Недоліки – обмежена ефективність у виявленні полісемії та використанні у спеціалізованих та доменних контекстах; обмеження у врахуванні синтаксичних та граматичних відношень між словами. Topic Modelling – статистичний метод, який виявляє приховані теми. Переваги – допомагає зрозуміти основні теми або дискусії в текстовому корпусі. Недоліки – необхідність правильного підбору гіперпараметрів; неоднозначність у визначенні тематичних відношень у тексті; інтерпретація тем може бути складною та вимагати експертного аналізу для визначення їх семантики [2, 3].

Підсумовуючи, можна зробити висновок, що у сучасному інформаційному світі автоматизоване опрацювання текстів відіграє ключову роль у багатьох галузях, включаючи пошукові системи, системи машинного перекладу та інше. Завдяки автоматизованій обробці можливо полегшити процеси пошуку інформації, аналізу даних, вилучення значущих фактів і надання релевантної інформації, а також скоротити час і зусилля, що витрачаються на опрацювання великих обсягів текстів, що сприяє підвищенню ефективності та якості роботи. Однак, ефективне опрацювання залишається викликом і потребує постійної уваги і досліджень з метою усунення недоліків і помилок [4].

Список літератури:

1. *Dashenkov, D.* Dataset for NLP-enhanced image classification/ *D. Dashenkov, K. Smelyakov, N. Sharonova* // COLINS (2). – 2023. – №3396 – С. 88 – 101.

2. *Egger, R.* Natural Language Processing (NLP): An Introduction: Making Sense of Textual Data/ *R. Egger, E. Gokce* // Springer. – 2022. – №7 – С. 307 – 334.

3. Future Processing NLP techniques: key methods that will improve your analysis [Електрон. ресурс]. – Режим доступу: http://surl.li/xwyzsl – NLP techniques: key methods that will improve your analysis.

4. *Khairova, N.* Models for effective categorization and classification of texts into specific thematic groups (using gender and criminal themes as examples)/ *N. Khairova, Y. Kupriianov, A. Vorzhevitina, O. Shanidze* // CEUR-WS. – 2024. – №4 – С. 37 – 49.