DOI 10.34185/1562-9945-5-154-2024-05
УДК 004.91

І.С. Макаров, Т.А. Ліхоузова

ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ДЛЯ АНАЛІЗУ ТЕКСТОВОЇ ІНФОРМАЦІЇ З TELEGRAM

Анотація. У сучасному світі інформаційних технологій розвиток інтернету призвів до стрімкого зростання кількості інформації. Відповідно, сьогодні особливої актуальності набувають застосунки, що здатні полегшити роботу з цією інформацією. Особливу увагу в цьому контексті заслуговують системи агрегації та класифікації текстової інформації, які використовуються для обробки даних з різноманітних джерел, зокрема з телеграм-каналів. Найбільш активною сферою досліджень є використання моделей машинного навчання для аналізу тексту, що відкриває нові можливості для підвищення ефективності обробки даних. Запропоновано програмне забезпечення – вебзастосунок для аналізу текстової інформації з Telegram. Можливі сфери застосування розробленого застосунку охоплюють широкий спектр галузей – від цифрового маркетингу та соціальних досліджень до аналізу новин та наукових досліджень.
Ключові слова: аналіз текстової інформації, моделі машинного навчання для аналізу тексту, BERT, класифікація.

Сучасні методи обробки текстових даних, зокрема машинне навчання, дозволяє нам визначати емоційне забарвлення тексту, класифікувати його за певними категоріями, виділяти основні теми та інші характеристики. Провідні наукові установи та ІТ-компанії активно досліджують цю галузь, розробляючи алгоритми для ефективної обробки та аналізу текстових даних. Зокрема, важливу роль відіграють роботи в галузі обробки природної мови (NLP) [1], що включають розробку алгоритмів для розуміння, інтерпретації та генерації людської мови машинами.

Дана робота зосереджена на створенні вебзастосунку, що спрямований на агрегацію та класифікацію текстової інформації з телеграм-каналів. Подібні застосунки відіграють ключову роль у структуруванні та систематизації текстових даних, що спрощує їх аналіз. У світлі стрімкого збільшення обсягів інформації в мережі Інтернет, таке програмне забезпечення виявляється актуальним як для сьогодення, так і у найближчому майбутньому.

Метою розробки є створення інструменту, що надасть користувачам можливість ефективно класифікувати, аналізувати та візуалізувати великі обсяги текстової інформації.

Аналіз останніх досліджень і публікацій. В контексті обробки текстових даних варто пам’ятати, що текстова інформація в цифровому форматі – це не тільки електронні книги чи статті, але й пости в соціальних мережах, повідомлення в месенджерах та блоги. В наш час соціальні мережі та месенджери почали відігравати ключову роль в інформуванні суспільст-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

ва та формуванні думок мільйонів українців. Так, за даними Київського міжнародного інституту соціології найпопулярніше джерело інформації для громадян нашої країни є телеграм-канали [2].

Популярність та актуальність задач агрегації та аналізу текстової інформації призвела до появи та розвитку великої кількості програмних продуктів у даній сфері. Однак, розгляд існуючих застосунків виявляє ряд проблем, які суттєво обмежують їхній потенціал та ефективність використання в академічних та дослідницьких цілях. Основні моменти, на яких буде побудоване порівняння, це чи існує можливість адаптувати роботу застосунку під конкретну задачу користувача, чи доступні користувачу інструменти для проведення досліджень у сфері соціальних наук, наявність алгоритмічної прозорості та доступ до вихідного коду застосунків для спільноти. Окрім цього ми звернемо увагу на доступний функціонал: можливість експорту сирих даних, що були використанні у побудові графіків, що є корисним для будь-якого дослідника, можливість побудови «хмари слів».

Таблиця 1

Порівняння з аналогами (Kin-TXT – розроблене)

Функціонал	Kin-TXT	Revuze	Talk Walker	Moneky Learn	Пояснення
Експорт сирих даних	+	-	-	+	Можливість зберегти сирі дані для інших досліджень.
Інтеграція користувацьких класичних моделей машинного навчання	+	-	-	+	Дозволяє використовувати користувацькі моделі, засновані на традиційних алгоритмах ML.
Інтеграція користувацьких нейронних мереж	+	-	-	-	Можливість використання користувацьких нейронних мереж.
Побудова хмари слів	+	-	+	+	Візуалізація найбільш часто вживаних слів у датасеті для швидкого ідентифікування ключових тем.
Вбудовані моделі для досліджень у сфері соц. наук	+	-	+	+	Наявність готових моделей ML для задач, що корисні для досліджень у соціальних науках.
Алгоритмічна прозорість	+	-	-	+	Відкритість деталей алгоритмів, що лежать в основі обробки та аналізу даних.
Відкритість продукту для спільноти	+	-	-	-	Можливість учасників спільноти вносити зміни або адаптувати продукт під власні потреби.

Результати та основний матеріал дослідження. В рамках даної роботи передбачалась розробка моделі для категоризації новинних повідомлень за тематикою. Це вимагає застосування ефективних алгоритмів машинного навчання, здатних точно ідентифікувати ключові особливості тексту та відносити його до відповідної категорій. Декілька популярних алгоритмів для класифікації текстів включають: наївний баєсів класифікатор [3], метод опорних векторів [4], RNN [5], BERT [6]. Кожен з цих алгоритмів має свої особливості.

Проаналізувавши всі переваги та недоліки кожного з алгоритмів текстової класифікації було прийняте рішення про використання нейронної мережі архітектури BERT для задачі класифікації новин. Обмежена кількість даних, що були зібрані та розмічені вручну в рамках даної роботи, виключає можливість тренування будь-якої нейронної мережі з нуля. Окрім цього, функціонал застосунку не передбачає класифікацію даних в режимі реального часу, відповідно обмежена швидкість роботи мережі на основі BERT не буде критичним недоліком. На додаток у вільному доступі існує велика кількість різноманітних перед тренованих моделей даної архітектури, навчених на великих корпусах даних. Використання такого підходу з переднавченою та в подальшому адаптованою моделлю дозволяє створити ефективну класифікаційну систему навіть маючи обмежений набір даних. Використано базову multilingual версію моделі BERT. Дана модель підтримує обробку тексту одразу на кількох мовах, включаючи українську та російську. Це критично важливо, оскільки завданням застосунку є аналізувати українські новинні потоки (як україномовні так і російськомовні) та російськомовну пропаганду. Можливість моделі ефективно працювати з текстами на обох мовах є ключовою для досягнення поставлених цілей.

Головною функцією програмного забезпечення є завантаження користувацьких моделей машинного навчання, побудова статистичних звітів на основі вбудованої моделі, побудова статистичних звітів на основі користувацької моделі, побудова звітів типу «хмара слів» на основі вбудованої моделі та побудова звітів типу «хмара слів» на основі користувацької моделі, більше функцій можна побачити на рисунку 1.

Ретельний аналіз варіантів використання дозволив визначити ключові сценарії взаємодії користувачів із системою, виокремивши основні функції, які має підтримувати розроблюваний застосунок. Вивчення системних вимог дало змогу уточнити технічні аспекти та умови експлуатації системи, що є важливим для забезпечення її стабільної роботи та інтеграції з іншими сервісами. Формулювання функціональних вимог відобразило детальний опис основних задач, які повинна вирішувати система, включаючи підтримку користувацьких моделей, управління шаблонами візуалізації, генерацію та управління звітами. Це дозволило чітко визначити обсяг робіт та функціональні можливості застосунку. Нефункціональні вимоги, в свою чергу, окреслили ключові якісні характеристики системи, такі як масштабованість, надійність, точність, юзабіліті, безпека, що забезпечують зручність використання застосунку та його ефективність.

Для вебзастосунку обрано мікросервісну архітектуру. Такий вибір обумовлений потребою у високій масштабованості, гнучкості управління окремими компонентами системи, спрощенням процесу розробки та тестування за допомогою декомпозиції функціоналу на незалежні сервіси.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Вибір засобів розробки був обґрунтований на основі аналізу наявних інструментів та специфіки задач, що стоять при побудові даного проєкту. Зокрема, було визначено використання мови програмування Python з використанням фреймворку FastAPI для бекенду та React для фронтенду, забезпечуючи таким чином ефективність та сучасність вебзастосунку.

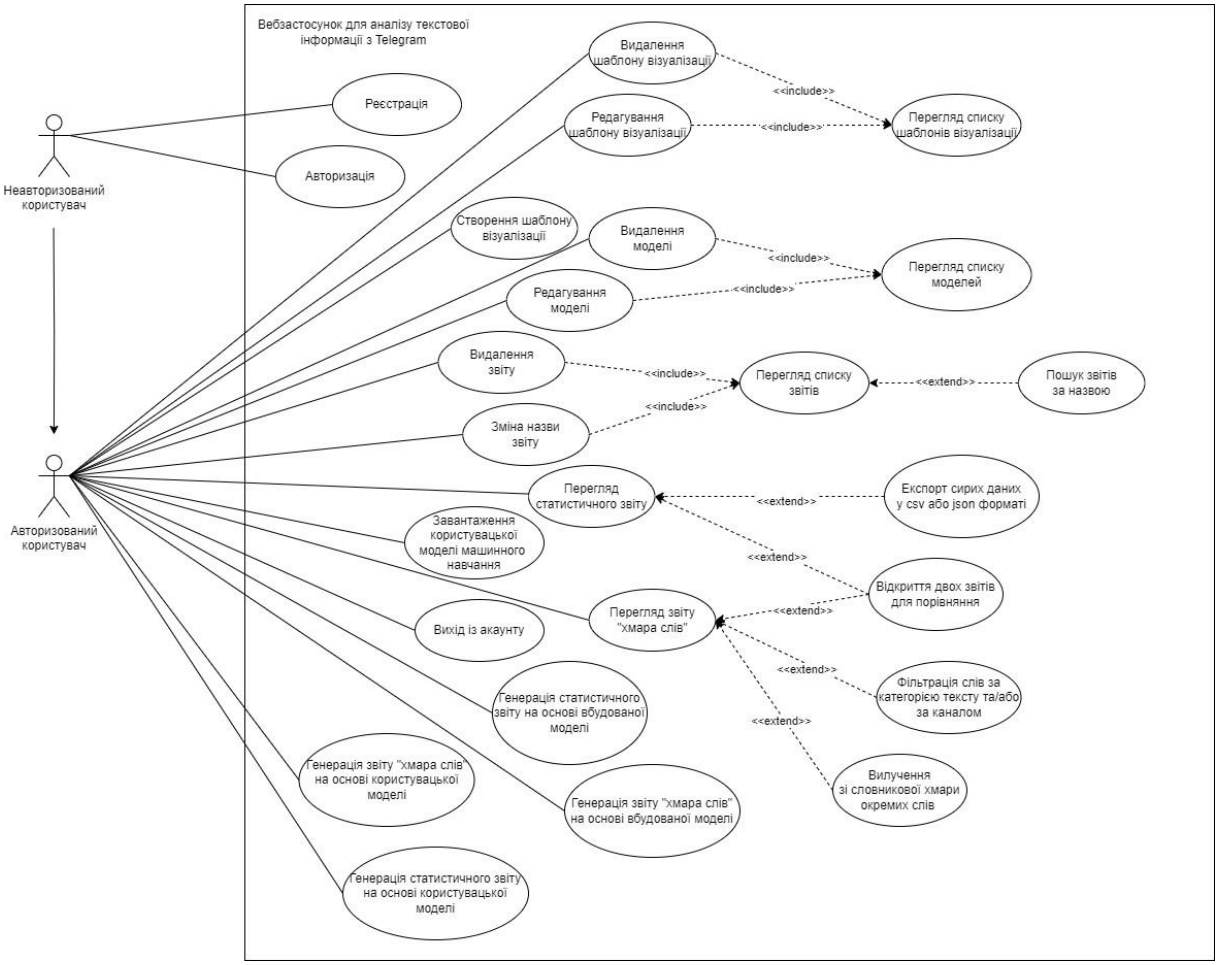

Рисунок 1 – Діаграма варіантів використання

В якості фундаменту для класифікаційної моделі обрано Base Multilingual BERT (mBERT), який забезпечує широкі можливості для розуміння різноманітних мов. Дана модель має використовує 12 шарів енкодерів та має близько 110 мільйонів параметрів.

Поверх базової мультимовної версії BERT було додано два шари нейронів, задачею яких буде власне класифікація новин. Перший додатковий шар складається з 128 нейронів і відповідає за виділення більш специфічних особливостей із векторних представлень mBERT, які можуть бути важливими для розрізнення категорій новин.

Другий шар, який є кінцевим у моделі класифікації, містить 5 нейронів, що відповідають за 5 категорій новин: Crisis, Economical, Political, Corruption та Other. Кожен нейрон у цьому шарі представляє одну категорію, і активація цих нейронів відбувається через функцію softmax, яка перетворює вихідні значення нейронів на ймовірності приналежності введення до кожної з категорій.

Для побудови моделі було зібрано та розмічено близько 2500 новинних постів різних категорій. Кожен пост було перечитано та вручну анотовано з метою забезпечення якісних тренувальних даних.

Зібрані дані були розділені на навчальну та тестову вибірку у відношенні 80 до 20. Процес навчання моделі проходив протягом п'яти епох, що означає, що весь набір даних був використаний для тренування п'ять разів. В якості оптимізаційного алгоритму було обрано Adam [7], який є одним з найпопулярніших та ефективних алгоритмів для тренування глибинних нейронних мереж. Adam оптимізує процес навчання за допомогою адаптації швидкості навчання для кожного параметра моделі, що дозволяє досягти кращих результатів за меншу кількість ітерацій.

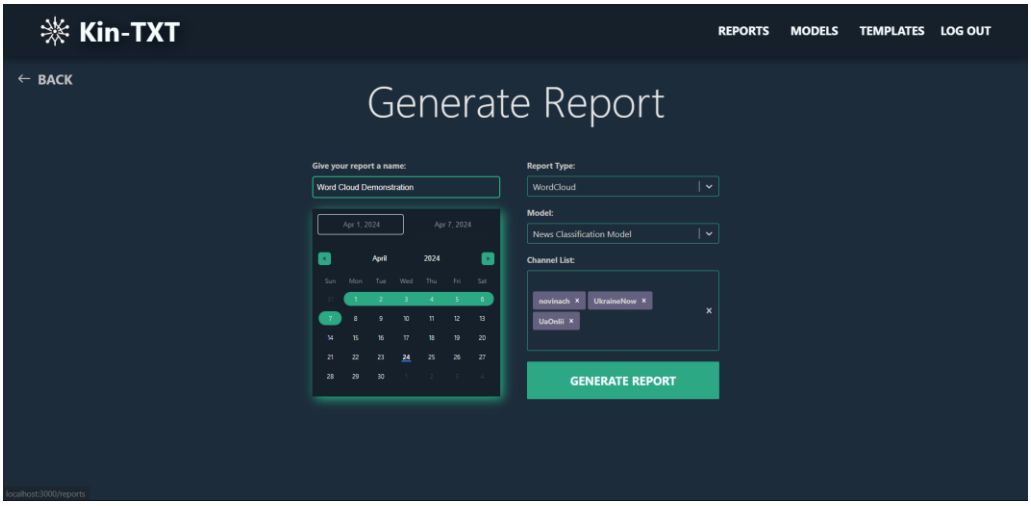

Рисунок 2 – Приклад форми для генерації звіту типу «хмара слів»

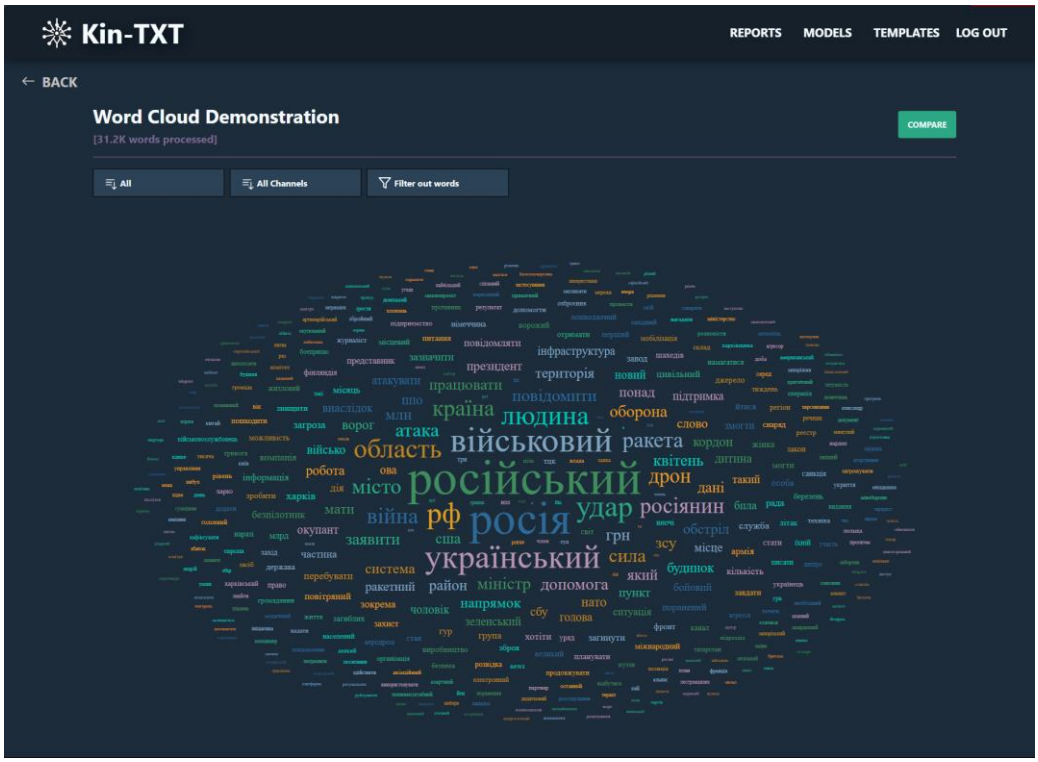

Рисунок 3 – Приклад відображення звіту типу «хмара слів»

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

По завершенню навчання модель була протестована на окремому тестовому наборі даних, що не входив до тренувального датасету, демонструючи точність класифікації на рівні 97.8%. Такий високий показник точності свідчить про те, що модель добре узагальнила навчальні дані та здатна ефективно класифікувати новини за категоріями, що є підтвердженням досягнення мети розробки.

Висновки. Розроблений вебзастосунок вносить свій вклад у сферу агрегації, обробки та візуалізації текстових даних. Враховуючи відкритість даного проєкту для спільноти та актуальність аналізу соціальних тенденцій, дана робота може сприяти розвитку наукових досліджень в області соціології та політології, де аналіз соціальних мереж і медіа є ключовим для розуміння суспільних процесів.

Застосунок дозволяє науковцям, аналітикам та студентам не тільки доступатися до потужних інструментів для обробки великих обсягів даних, але й участь у його вдосконаленні та налаштуванні під конкретні дослідницькі задачі завдяки його відкритому коду. Це сприяє не лише академічній співпраці, але й стимулює інноваційні підходи та методики в аналізі текстової інформації, роблячи дослідження більш глибокими та об'єктивними.

ЛІТЕРАТУРА / REFERENCES

1. Що таке NLP [Електронний ресурс] // Metinvest Digital. – Режим доступу до ресурсу: https://metinvest.digital/ua/page/1052. - Назва з екрана.

2. Результати всеукраїнського опитування для Консультативної місії Європейського Союзу в Україні [Електронний ресурс] / Київський міжнародний інститут соціології. - 2023. - Режим доступу:
https://kiis.com.ua/materials/pr/20231026_r/AReport_PublicSurvey_EUAM_sept2023_ukr_public.pdf. - Назва з екрана.

3. Naïve Bayes [Електронний ресурс]. – Режим доступу до ресурсу:
https://www.ibm.com/topics/naive-bayes. - Назва з екрана.

4. Cortes C., Vapnik V. Support-Vector Networks [Електронний ресурс] // Machine Learning. - 1995. - Vol. 20, No. 3. - Pp. 273-297. - Режим доступу:
https://link.springer.com/article/10.1007/BF00994018. - Назва з екрана.

5. Goodfellow I. Deel Learning / Goodfellow I. Bengio Y. Courville A. – Cambridge, MA : MIT Press, 2016. – 367 с.

6. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Електронний ресурс] / Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova // arXiv preprint arXiv:1810.04805. - 2018. - Режим доступу: https://arxiv.org/abs/1810.04805. - Назва з екрана.

7. Kingma D., Ba J. Adam: A method for stochastic optimization [Електронний ресурс] / Diederik P. Kingma, Jimmy Ba // arXiv preprint arXiv:1412.6980. - 2014. - Режим доступу:
https://arxiv.org/abs/1412.6980. - Назва з екрана.

Received 14.05.2024.
Accepted 16.05.2024.

Software for analyzing text information from Telegram

In the modern world of information technologies, the development of the Internet has led to a rapid increase in the amount of information. Accordingly, applications that can facilitate work with this information are gaining special relevance today. In this context, systems of aggregation and classification of textual information, which are used to process data from various sources, including telegram channels, deserve special attention.

World trends in this area indicate a growing need for improving tools for processing textual information, which stimulates scientific research and the development of new technologies. The importance of such systems is confirmed by active developments in this field by IT companies and universities around the world. The most active field of research is the use of machine learning models for text analysis, which opens up new opportunities for increasing the efficiency of data processing.

In the context of developing systems for the analysis of textual information, many existing solutions face challenges related to scalability and adaptation to various types of data. However, this work seeks to approach the development of such software from a different angle, focusing its attention on the flexibility and openness of the system to the community. The application supports a limited set of built-in machine learning models optimized for different text data classification tasks, while offering users the ability to integrate their own models according to their unique needs. This approach not only provides a foundation for a wide range of applications, but also promotes community development and innovation by taking advantage of collective intelligence.

Software is offered - a web application for analyzing text information from Telegram. Possible areas of application of the developed application cover a wide range of industries - from digital marketing and social research to news analysis and scientific research.

Keywords: analysis of text information, machine learning models for text analysis, BERT, classification.

Макаров Ілля Сергійович – студент кафедри інформатики та програмної інженерії, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Ліхоузова Тетяна Анатоліївна – к.т.н., доцент кафедри інформатики та програмної інженерії, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського».

Makarov Illia – student, Department of Informatics and Software Engineering National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».

Likhouzova Tetiana – PhD, associate professor, Department of Informatics and Software Engineering National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute».

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)