Список використаних джерел

1. Глушков В. М. Макроэкономические модели и принципы построения ОГАС. – М. : Статистика, 1975. – 160 с.

2. Глушков В. М. Мышление и кибернетика. – М. : Знание, 1966. – 36 с.

3. Ярошевский М. Г. Кибернетика – «наука» мракобесов // Литературная газета. – 5 апреля 1952 г.

4. Babbage Ch. On the Economy of Machinery and Manufactures. 4th ed. – London : John Murray, 1846. – 408 p.

5. Gorz A. Écologica. – Paris : Galilée, 2008. – 168 p.

6. Marx K. Das Elend der Philosophie // Marx K., Engels F. Werke. – Berlin: Dietz Verlag, 1972. – Bd. 4. – S. 63–182.

7. Sen A. On Ethics and Economics. – Oxford : Blackwell Publishing, 1988. – XIII, 133 p.

Молчанова М.О., Залуцька О.О., Бармак О.В.,

м. Хмельницький

m.o.molchanova@gmail.com

МЕТОД ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ТОНАЛЬНОСТІ ТЕКСТІВ

Емоційна тональність тексту вказує на емоційний характер або емоційне забарвлення текстів. Ця характеристика визначає, які емоції чи почуття виражені в тексті, чи є вони позитивними, негативними або нейтральними. Тональність тексту може бути важливою для багатьох застосувань, таких як аналіз відгуків користувачів, визначення настроїв ринку, виявлення відгуків у соціальних мережах, фільтрація інформації, для покращення якості комунікації та взаємодії з користувачами в різних додатках і системах. Отож, метою роботи є розробка методу інтелектуального аналізу тональності текстів.

Задача автоматичного інтелектуального аналізу емоційної тональності текстів для визначення поведінкових намірів їх авторів зводиться до задачі класифікації: позитивна тональність (вказує на наявність позитивних емоцій у тексті) і негативна тональність (вказує на наявність негативних емоцій).

В межах даного дослідження, оцінка емоційної тональності текстів виконувалась відносно відгуків у засобах електронної комерції [1]. У свою

чергу, відгуки електронної комерції мають наступні особливості: обмежений обсяг контенту (до 500 слів); малий обсяг контенту (1-3 слова); використання суржику, професіоналізмів, жаргонів та інтегрованого мультимовного контенту.

Переважна більшість відгуків не перевищує 100 слів, а більш довгими, як правило, є негативні відгуки. В якості датасету було використано вибірку відгуків з платформи «Hotline» (https://hotline.ua). Такий вибір експериментальних даних обумовлено тим, що цікавить саме розмовний україномовний контент, який до того ж повинен бути розміченим. Оцінками слугують оцінки клієнтів, які залишають відгуки, де оцінка «Не рекомендую» – негативні відгуки, а «Рекомендую» – позитивні. Для видобутку відгуків було створене відповідне програмне забезпечення на базі бібліотеки Crawlee, та в подальшому оброблені створеним програмним застосунком на мові C#, розподілені на вибірки – «позитив» та «негатив». Загалом датасет складається із 7656 документів, де в навчальній вибірці знаходиться 6655 документів, із яких 1331 документ використано для валідації.

Для бінарної класифікації настроїв україномовних відгуків електронної комерції розглядалися BERT-подібні нейромережі, оскільки відомі дослідження свідчать, що словникові інструменти для вилучення настроїв із текстових даних мають явну перевагу з точки зору інтерпретації, але явно втрачають в точності. Авторами з результатом досліджень встановлено, що ukr-RoBERTa, ukr-ELECTRA та XLM-R large мають тенденцію демонструвати найвищу продуктивність, хоча XLM-R large та ukr-ELECTRA мають тенденцію працювати краще на довших текстах, тоді як ukr-RoBERTa значно перевершує інші моделі на коротших послідовностях [2]. Оскільки дослідження проводиться на текстах відгуків інтернет-платформи «Hotline», які, як правило, є короткими текстовими повідомленнями, тому було прийнято рішення використовувати нейромережу RoBERTa.

Конфігурація нейронної мережі для інтелектуального аналізу емоційної тональності текстової інформації на базі обраного датасету та типу нейромережі має структуру, показану на Рисунку 1. Так, у вхідному шарі відбувається перетворення вхідної текстової інформації на тензор Keras, тобто символічний тензороподібний об’єкт, що доповнюється атрибутами, які дозволяють сформувати модель Keras за вхідними й вихідними даними моделі. Надалі

тензор подається на вхід шару попередньої обробки, яка включає в себе обгортку викликаного об'єкта для використання як шару Keras на базі попередньо навченої моделі попередньої обробки тексту. Дана модель використовує SentencepieceTokenizer, що токенізує тензор рядків UTF-8 та є неконтрольованим токенізатором і детокенізатором тексту.

Наступним шаром є RoBERTa енкодер, який працює на основі попередньо навченої моделі «xlm_roberta_multi_cased_L-12_H-768_A-12», що є результатом неконтрольованого крос-мовного репрезентативного навчання в масштабі (XLM-RoBERTa) [3], та попередньо навчена на 2,5 ТБ відфільтрованих даних CommonCrawl, що містять 100 мов. Після чого шар dropout випадково встановлює одиниці введення на 0 із частотою швидкості на кожному кроці під час навчання, що допомагає запобігти перенавчанню. Вхідні дані, для яких не встановлено значення 0, масштабуються таким чином, щоб сума всіх вхідних даних не змінювалася. Останнім кроком в моделі є власне класифікація тональності, що здійснюється з використанням функції Dense та видає результат від 0 до 1, де 0 – негативний відгук, а 1 – позитивний відгук.

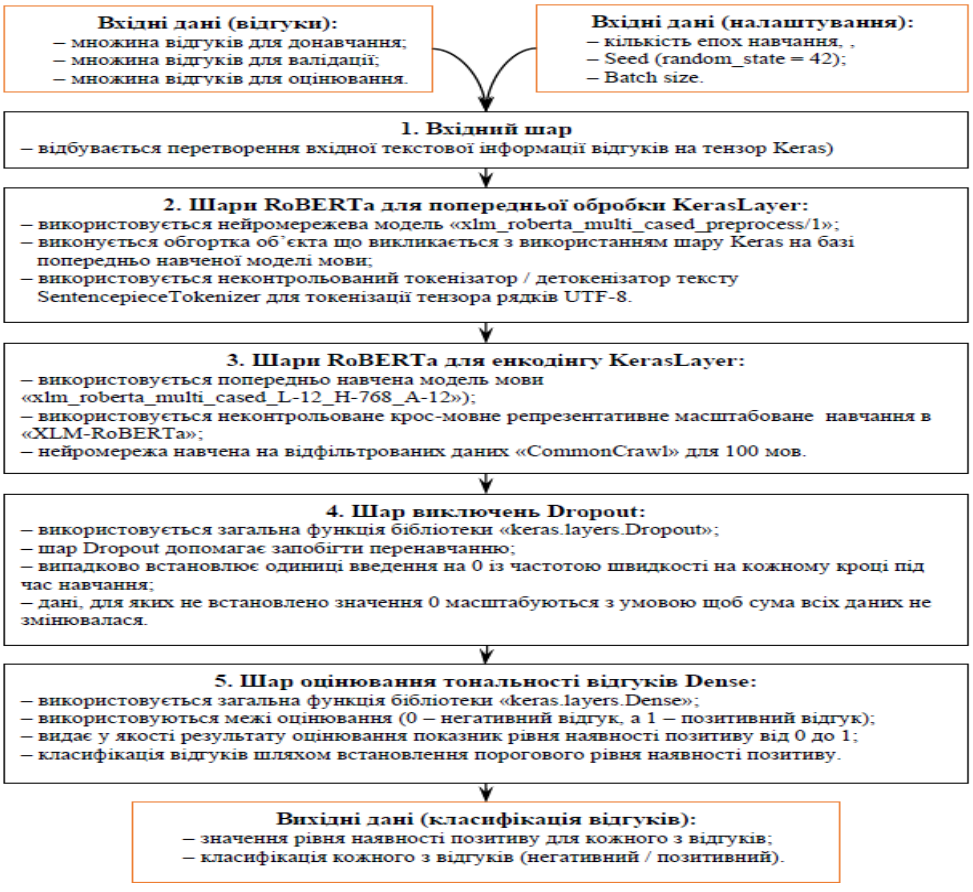

Рис. 1. Схема роботи класифікатора на основі нейромережі RoBERTa для інтелектуального аналізу тональності текстів

Далі запропонована модель проходить донавчання вищеописаною вибіркою. Донавчання проводилось із різною комбінацією кількісних показників параметрів, таких як: кількість епох навчання, Seed, Batch size.

Дослідження відгуків, яких немає у навчальній та тестовій вибірках, показало високу ефективність пропонованої архітектури. Отримані результати свідчать, що при використанні вибірки для валідації точність класифікації не росте. А функція втрат після 3-ї ітерації для вибірки для валідації мала тенденцію до незначного зростання.

Отже, було розглянуто метод інтелектуального аналізу тональності текстів з використанням нейронної мережі RoBERTa та супутні питання: формування розміченого датасету для навчання нейромережі, підбор та налаштування нейромережевого класифікатора, побудову семантичної моделі мови. В результаті, для комбінованих мультимовних відгуків вдалося отримати точність 0.92, в той час як функція втрат мала значення 0.29.

Розроблений метод має особливості застосування, зокрема, його доцільно використовувати до визначення тональності саме коротких текстів (довжиною до 500 слів) на українській мові, що можуть містити суржик та іншомовні вкладення слів.

Список використаних джерел

1. Zalutska O., Molchanova M., Sobko O., Mazurets O., Pasichnyk O., Barmak O., Krak I. Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. CEUR Workshop Proceedings, 2023, vol. 3387, pp. 561–571.

2. RoBERTa: An optimized method for pretraining self-supervised NLP systems URL: https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems.

3. XLM-RoBERTa (base-sized model). URL: https://huggingface.co/xlm-roberta-base.