УДК 004.4

Залуцька О.О., Молчанова М.О., Віт Р.В., Мазурець О.В.

Хмельницький національний університет

КОНФІГУРУВАННЯ НЕЙРОННОЇ МЕРЕЖІ ДЛЯ КЛАСИФІКАЦІЇ ЕМОЦІЙНОЇ ТОНАЛЬНОСТІ ТЕКСТОВОЇ ІНФОРМАЦІЇ ЗА ПОКАЗНИКАМИ СЕМАНТИЧНОЇ ЗВ'ЯЗНОСТІ

Описано результати досліджень з конфігурування нейронної мережі класу трансформер для класифікації емоційної тональності текстової інформації за показниками семантичної зв'язності. Обґрунтовано вибір оптимальних параметрів, необхідних для ефективної класифікації емоційної тональності текстів.

The results of research on the configuration of a transformer-class neural network for the classification of the emotional tonality of text information based on indicators of semantic connectivity are described. The choice of optimal parameters necessary for effective classification of the emotional tonality of texts is substantiated.

Класифікація емоційної тональності текстової інформації є методом вилучення та розпізнавання оцінок користувачів щодо продуктів і моделей та має різні підходи з використанням алгоритмів машинного навчання для класифікації емоцій, що стоять за цим текстом [1]. До прикладу, аналіз настроїв твітів для розуміння сприйняття людьми певних новин, оцінка взаємодії людини з роботом, формування системи рекомендації у виборі товарів тощо.

Розв'язання завдання класифікації емоційної тональності україномовних текстів на прикладі відгуків сервісів електронної комерції може застосовуватись як для розуміння сприйняття людьми певних новин, так і для комерційних цілей як то оцінки роботи менеджера тощо.

У напрямку класифікації емоційної тональності текстової інформації більшість публікацій присвячено саме роботі з англомовними текстами, оскільки є достатня кількість розмічених наборів даних, на кшталт IMDB (набір розмічених даних, що містить понад 50 000 оглядів фільмів) та набір розмічених за емоційним забарвленням відгуків з інтернет-магазину «Amazon». Що ж стосується досліджень української мови, перша проблема з якою зіштовхуються науковці, стосується експериментальних даних [2]. В основному науковці такі дані збирають самі, що є трудомістким процесом, та зазвичай такі дані не є розміченими, їх потрібно розмічувати «вручну».

Для класифікації емоційної тональності текстової інформації використано варіацію нейронної мережі RoBERTa (скорочення від «Надійно оптимізований підхід BERT»), яка є варіантом моделі BERT (Bidirectional Encoder Representations

from Transformers), яку розробили дослідники Facebook AI [3]. Як і BERT, RoBERTa є мовною моделлю на основі трансформера, яка використовує самоувагу для обробки вхідних послідовностей і створення контекстуалізованих представлень слів у реченні.

Однією з ключових відмінностей між RoBERTa та BERT є те, що RoBERTa навчався на значно більшому наборі даних і з використанням ефективнішої процедури навчання. Під час навчання RoBERTa використовує техніку динамічного маскування, що допомагає моделі вивчати більш надійні та узагальнені представлення слів.

Так як класифікація емоційної тональності текстової інформації за показниками семантичної зв'язності на основі нейромережевого підходу є сьогодні актуальним напрямом наукових досліджень, для української мови на сьогодні також є деякі напрацювання. Одними з яких є попередньо навчена мультимовна модель препроцесингу, що працює також і з українською мовою та ще з понад 50 іншими мовами [4], та входить до складу моделей бібліотеки Tensorflow_hub мови Python. На базі цих моделей пропонується створити модель, що буде донавчено на вищеописаній вибірці експериментальних даних. Вибір мультимовних моделей обумовлено тим, що як вже було вище наведено, тексти можуть містити текст не тільки літературною українською мовою.

Конфігурація нейронної мережі для класифікації емоційної тональності текстової інформації на базі обраного типу нейромережі має наступну структуру. На вхідному шарі відбувається перетворення вхідної текстової інформації на тензор Keras, тобто символічний тензороподібний об'єкт, який доповнюється атрибутами, які дозволяють побудувати модель Keras за вхідним та вихідними даними моделі. Надалі тензор подається на вхід шару попередньої обробки, яка включає в себе обгортку об'єкта, що викликається, для використання як шару Keras на базі попередньо навченої моделі попередньої обробки тексту [4]. Дана модель використовує SentencepieceTokenizer [5], що токенізує тензор рядків UTF-8 та є неконтрольованим токенізатором і детокенізатором тексту.

Наступним шаром є RoBERTa енкодер. Цей шар працює на основі попередньо навченої моделі «xlm_roberta_multi_cased_L-12_H-768_A-12» [6], що є результатом неконтрольованого крос-мовного репрезентативного навчання в масштабі (XLM-RoBERTa), та попередньо навчена на 2,5 ТБ відфільтрованих даних CommonCrawl, що містять 100 мов.

Наступним шаром є шар dropout, що 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 безпосередньо класифікація, що здійснюється з використанням функції Dense та видає результат від 0 до 1, що є мірою позитиву в україномовних відгуках електронної комерції. Де 0 – негативний текст, а 1 – позитивний текст.

Далі запропонована модель проходить донавчання під вищеописану вибірку. Донавчання проводилось із різною комбінацією кількісних показників параметрів, таких як: кількість епох навчання, Seed, Batch size [7].

Кількість епох навчання показує, скільки разів модель підлягає навчанню. Параметр Seed буде взято 42, з огляду на те, якщо не встановити для random_state значення 42, щоразу, коли знову буде запускатись програмний код, він створюватиме інший тестовий набір. Batch size – кількість навчальних прикладів, що використовуються в межах однієї ітерації. Дуже важко відразу визначити, який ідеальний розмір партії для потреб конкретної задачі, тому даний параметр буде підібрано експериментальним шляхом.

Відповідно до обраних параметрів, визначались показники оцінки функціональності моделі класифікації емоційної тональності текстової інформації, такі як: час навчання в секундах, точність та втрати. У якості функції втрат використовувалась бінарна крос-ентропічна функція. Точність для проведеного дослідження визначається як ділення кількості правильних відповідей на загальну кількість відповідей.

Враховувались одержані показники оцінки функціональності (час навчання, точність та втрати) різних параметрів моделей налаштування (кількість епох навчання, seed, batch size) нейромережевого класифікатора. Оскільки досліджувана версія RoBERTa є мультимовним трансформером, донавченим на білінгвістичних даних, в цілому нейромережа не має проблем з ідентифікацією настроїв.

При дослідженні текстів, яких немає в навчальній та тестовій вибірках показано високу ефективність пропонованої архітектури. Навчальна вибірка не чистилась «вручну», тому допускається, що може бути певний відсоток хибно-класифікованих текстів, проте це не дає значного впливу на кінцеву точність класифікації емоційної тональності текстової інформації, що написані не лише чистою українською мовою, а й містять суржик та білінгвістичні дані. На Рисунку 1 графік ілюструє зміни параметра точності в залежності від пройдених епох, а Рисунку 2 – зміни функції втрат для комбінації параметрів навчання: 3 епохи, 64 розмір батча.

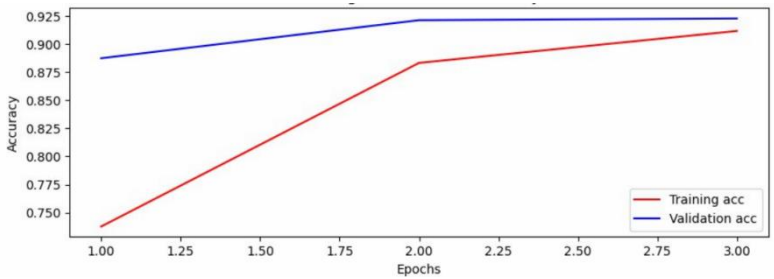

Рисунок 1 – Ілюстрація процесу навчання за епохами за показником точності за кількості епох навчання 3

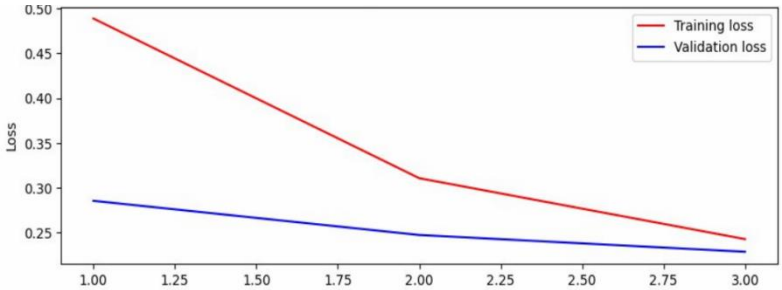

Рисунок 2 – Ілюстрація процесу навчання за епохами за показником функції втрат за кількості епох навчання 3

Графік на Рисунку 1 свідчить про недостатню кількість епох навчання для стабілізації результату, оскільки показник Accuracy мав тенденцію до зростання, а показник функції втрат – до спадання, не застигши на одному рівні.

Проте, продовживши експеримент, змінивши кількість епох навчання до 10, були отримані результати, проілюстровані на Рисунках 3 та 4.

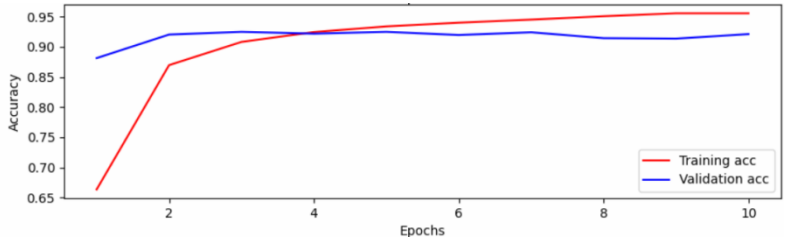

Рисунок 3 – Ілюстрація процесу навчання за епохами за показником точності за кількості епох навчання 10

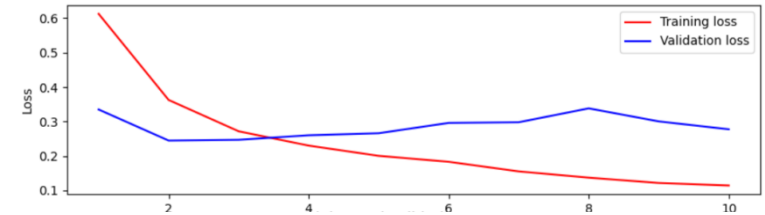

Рисунок 4 – Ілюстрація процесу навчання за епохами за показником функції втрат за кількості епох навчання 10

Отримані результати свідчать, що при використанні вибірки для валідації точність класифікації не росте. А функція втрат взагалі після 3ї ітерації для вибірки

для валідації мала тенденцію до незначного зростання. Проте, такі результати можуть свідчити про те, що вибірки недостатньо відфільтровані. Оскільки перевірка нейромережі на текстах, що не містяться в базі дала практично безпомилкові результати для 40 текстів, які дійсно містили емоцію. Графік ілюстрації проходження процесу донавчання по епохам показано на Рисунках 5 та 6.

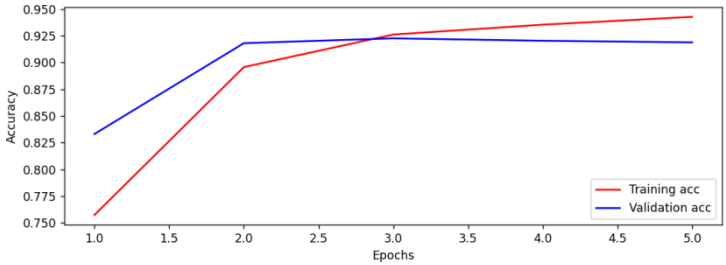

Рисунок 5 – Ілюстрація процесу навчання за епохами за показником точності з донавчанням

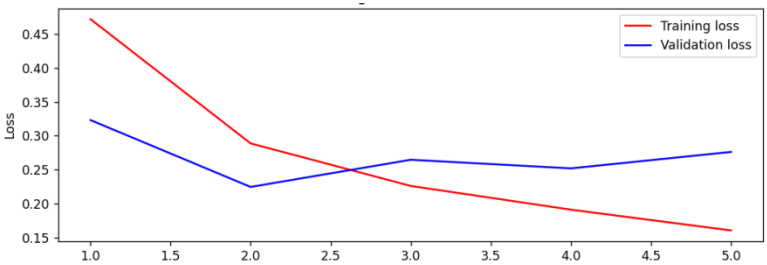

Рисунок 6 – Ілюстрація процесу навчання за епохами за показником функції втрат з донавчанням

Результати даного експерименту з класифікації емоційної тональності текстової інформації свідчать про те, що вибірка як вже було вище сказано, не була очищена вручну. Тому, зі зростом кількості епох нейромережа починає просто «запам'ятовувати», які тексти куди належать, про що свідчить червона лінія на графіках 3-4 та 5-6. Так як для навчальної вибірки функція втрат значно менша, а точність – значно вища. Проте, отримані показники функції втрат та точності пов'язані з тим, що вибірка не була відфільтрована «вручну», і містила тексти, які включали беземоційні коментарі, часто з одного слова або фрази. До того ж, проведений аналіз оцінки тональності показав на практиці з 40 фраз, яких не має ні в навчальній, ні в тестовій вибірках, і які були попередньо оцінені експертом, що нейронна мережа справляється з завданням безпомилково, при чому тексти містили як стилістичні, так і орфографічні помилки та були представлені мультимовними даними.

Отже, було розглянуто сучасний стан напряму семантичної обробки тексту, а саме класифікації емоційної тональності текстової інформації. Однією із найбільш точних нейромереж визначили архітектуру BERT, проте для аналізу коротких документів краще себе показала її модифікація – RoBERTa. Для оцінки роботи запропонованої архітектури для класифікації емоційної тональності текстової інформації було використано точність та функцію втрат. Для комбінованих мультимовних текстів вдалося отримати точність 0.92, в той час як функція втрат мала значення 0.29.

Запропонований підхід до класифікації емоційної тональності текстової інформації має певні обмеження. Доцільно його застосовувати до визначення тональності коротких текстових текстів (довжиною до 500 слів), представлених на українській мові та можуть містити суржик та іншомовні вкладення слів. Зміна вмісту навчальної вибірки впливає на результат навчання нейронної мережі, і відповідно впливає на ефективність класифікації емоційної тональності текстової інформації. З часом в побутовій мові можуть відбуватися зміни, які також впливають на хід та результати класифікації емоційної тональності текстової інформації.

Перелік посилань

1. Mann, S., Arora, J., Bhatia, M., Sharma, R., Taragi, R, Twitter Sentiment Analysis Using Enhanced BERT, in: Kulkarni, A.J., Mirjalili, S., Udgata, S.K. Intelligent Systems and Applications. Lecture Notes in Electrical Engineering, vol 959. Springer, Singapore, 2023, pp. 263-271.
2. Panchenko, D., Maksymenko, D., Turuta, O., Yerokhin, A., Daniiel, Y., Turuta, O., Evaluation and Analysis of the NLP Model Zoo for Ukrainian Text Classification, in: Information and Communication Technologies in Education, Research, and Industrial Applications. ICTERI 2021, Communications in Computer and Information Science, vol 1698. Springer, Cham.
3. Ai.Facebook.Com., 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.
4. Tfhub.Dev., Text preprocessing model xlm_roberta_multi_cased_preprocess. URL: https://tfhub.dev/jeongukjae/xlm_roberta_multi_cased_preprocess/1.
5. Tensorflow.Org., Text.SentencepieceTokenizer. URL: https://www.tensorflow.org/text/api_docs/python/text/SentencepieceTokenizer.
6. Tfhub.Dev., Unsupervised Cross-lingual Representation Learning at Scale. xlm_roberta_multi_cased_L-12_H-768_A-12. URL: https://tfhub.dev/jeongukjae/xlm_roberta_multi_cased_L-12_H-768_A-12/1.
7. Залуцька О.О., Молчанова М.О., Мазурець О.В., Мельник О.І., Скрипник Т.К. Метод інтелектуального аналізу емоційної тональності текстової інформації для визначення поведінкових намірів нейромережевими засобами. Науковий журнал «Вісник Хмельницького національного університету» серія: Технічні науки. Хмельницький, 2023. №5 (325). Т.1. С. 67-73.