УДК 004.55

ДОСЛІДЖЕННЯ ТА РОЗРОБКА МОДЕЛІ НЕЙРОННОЇ МЕРЕЖІ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ СЕНТИМЕНТ АНАЛІЗУ ТЕКСТУ

Платоненко Сергій Дмитрович
к.т.н, доцент каф. ІС, Ніколенко Анатолій Олександрович
Національний університет «Одеська політехніка», Україна

АНОТАЦІЯ. Досліджено та розроблено модель нейронної мережі для вирішення задачі сентимент аналізу тексту на базі рекурентної нейронної мережі.

Вступ. Сентимент аналіз (СА) або аналіз тональності тексту – це завдання обробки природніх мов (ОПМ), яке має на меті вилучення, розпізнавання та класифікацію суб'єктивного ставлення автора тексту до певної теми на позитивні, негативні чи нейтральні категорії або ідентифікацію емоцій: злість, страх, передчуття, відраза, радість, печаль, здивування, довіра [1].

Основним джерелом СА є дані з соціальних мереж, месенджерів, коментарі до онлайн публікацій або відео блогів, результати он-лайн опитувань, дискусії певних новин, зібраних автоматично через Інтернет. Невпинний зріст кількості користувачів соціальних мереж перетворює соціальні мережі на середовище інформаційної взаємодії і все більше застосовується для реклами, пропаганди, формування точки зору та психологічних впливів. Наразі результати СА застосовуються для прогнозування політичних рейтингів, настроїв, намірів, уподобань людей, емоційних відгуків стосовно товарів. Для людини не виникає труднощів під час аналізу текстової інформації, навіть, якщо вона містить полісемію, омонімію, займенникову анафору чи еліпсис, а навчити комп'ютер розрізняти і правильно реагувати на такі ситуації доволі складно. Тому задачі ОПМ, і зокрема сентимент аналіз, є актуальними науково-практичними завданнями, які потребують подальшого розвитку і пошуку ефективних рішень [2].

Мета роботи – розробка та дослідження ефективної моделі нейронної мережі для вирішення задачі аналізу емоційного забарвлення тексту.

Основна частина роботи. Метод машинного навчання застосовується для аналізу настроїв шляхом класифікації тексту. Для цього створюється навчальний набір, в якому ручним чином позначається частина даних, та вивчається для побудови моделі класифікації. Ця модель використовується для класифікації та прогнозування тестових даних з невідомими тегами [3].

Один з ефективних підходів до класифікації тексту – застосування глибокого навчання, що базується на створенні нейронної мережі для класифікації тексту. У задачах аналізу настроїв найчастіше використовуються згорткові нейронні мережі (*CNN*) та повторювані нейронні мережі (*RNN*).

CNN оброблює текст у вигляді матриці, де кожен рядок - це векторний опис токену. Шари згортки оброблюють матрицю, використовуючи фільтри з фіксованою шириною та різними висотами. Після цього, шар субдискретизації зменшує розмірність карти ознак, щоб виділити найважливішу інформацію кожної згортки. Комбінація шарів згортки та субдискретизації дозволяє витягувати найбільш значущі n-грами з тексту.

Однак, фільтр *CNN* має обмежену ємність та не може вловити довгострокові залежності між словами, що не є сусідніми в реченні. Для цього можуть використовуватись мережі *RNN*, які моделюють послідовність з контекстним семантичним захопленням, що дозволяє застосувати вміст пам'яті до поточного сценарію. Проте традиційні нейронні мережі *RNN* можуть викликати проблему вибуху градієнту або його зникнення для довгих послідовностей даних.

Іншими словами, *CNN* корисний для виділення високорозмірних особливостей між локально сусідніми словами, тоді як *RNN* корисний для моделювання залежностей між несуміжними словами в послідовності.

LSTM (Long short time memory) є різновидом рекурентних нейронних мереж, яка здатна вивчати довготривалі залежності. Відмінність *LSTM* від стандартних *RNN* полягає у використанні "воріт" для регулювання кількості інформації, що проходить через осередок. Гейти складаються

з сигмовидного шару нейронної мережі і операції поточного множення, що дає можливість визначити, яка кількість інформації буде передана далі. Значення гейтів знаходяться в діапазоні від 0 до 1, де 0 означає, що інформація не буде передана, а 1 - що буде передана вся інформація.

LSTM мережі грають важливу роль у природній обробці мови завдяки їх потужності в обробці довгих текстів. Найбільш поширеними варіантами архітектури є рекурентна нейронна мережа з довгою короткочасною пам'яттю та керованим рекурентним блоком (*GRU*). Обчислення векторів прихованих станів у *RNN* виконується за допомогою керованого рекурентного нейрона, що дозволяє зберігати інформацію про віддалені залежності. При роботі методу зворотного поширення помилки, помилка переміщується по *RNN* від останнього кроку до першого. Однак, при досить малому початковому градієнті, наприклад менше 0,25, градієнт може майже повністю зникнути до третього або четвертого модуля, що може призвести до того, що приховані стани перших кроків не будуть оновлюватись [5].

Базові *LSTM* мережі сканують послідовності лише в одному напрямку, тоді як двонаправлена довготривала пам'ять (*BiLSTM*) [6] є подальшим вдосконаленням, що дозволяє сканування послідовності в обох напрямках, забезпечуючи одночасний доступ як до прямого, так і до зворотного контекстів.

BiLSTM мережа – це нейронна мережа, побудована на основі *LSTM*, яка може навчатися довготривалим залежностям без особливого навчання. *BiLSTM* складається з трьох шарів-фільтрів, що визначають стан комірок.

Архітектура *BiLSTM* передбачає врахування як наступного, так і попереднього контексту шляхом конкатенації їх. Спочатку обчислюється лівий контекст, потім правий контекст обчислюється у зворотному напрямку, а потім результати об'єднуються для створення повного представлення елемента вхідної послідовності.

Для оптимізації обчислення ймовірності класифікації може бути використаний шар умовно випадкових полів (*Conditional Random Field*) – дискримінаційна ймовірнісна модель, яка враховує контекст об'єкта, що класифікується, і використовується для прогнозування послідовностей.

Для використання умовно випадкових полів, спочатку визначаються необхідні функції, ініціалізуються ваги випадковими значеннями, а потім застосовується градієнтний спуск ітеративно до збіжності значень параметрів (у цьому випадку лямбда-значень). У порівнянні з іншими статистичними методами, метод *CRF* вимагає набагато менший обсяг навчальних даних, оскільки статистично значущі відношення можуть бути визначені як набір пов'язаних вершин [7].

Тому BiLSTM може вирішувати завдання моделі послідовності краще, ніж *LSTM*. Ці моделі нейронних мереж досягли значного успіху у завданні аналізу емоційної класифікації.

При роботі з нейронними мережами важливо мати можливість оцінити їх ефективність. Одним з показників ефективності може бути точність (*Precision*), формула 1 відображає відношення правильно визначених позитивних результатів до загальної кількості визначених позитивно результатів.

$$precision = \frac{\mathrm{TP}}{TP+FP}. \quad (1)$$

Повнота (*Recall*), формула 2 відображає відношення правильно визначених позитивних результатів до всіх позитивно відмічених.

$$recall = \frac{\mathrm{TP}}{TP+FN}. \quad (2)$$

Загальна точність (Accuracy), формула 3 відображає відношення кількості вірно визначених результатів до кількості усіх експериментів.

$$accuracy = \frac{\mathrm{TP+TN}}{TP+TN+FP+FN}. \quad (3)$$

Для оцінки ефективності моделей нейронних мереж було використано тестову вибірку коментарів до фільму з рівномірним розподіленням позитивних та негативних відгуків. Результати дослідження наведені в таблиці 1.

Таблиця 1 – Результати оцінки ефективності моделей

Model	Precision	Recall	Accuracy
CNN	0.8314	0.8072	0.8646
RNN	0.8358	0.8051	0.8640
LSTM	0.8416	0.8027	0.8650
BiLSTM	0.8474	0.8037	0.8666
BiLSTM-CRF	0.8529	0.8343	0.8816

Висновки. Досліджено різні моделі нейронних мереж для аналізу тональності текстів, використовуючи зібрані дані з різних джерел та публічних відгуків. Результати експериментів показали високу ефективність моделі *BiLSTM-CRF* у вирішенні задачі аналізу тональності. Для майбутньої роботи можна розглянути оцінку моделі на інших наборах даних, включно з більшими наборами даних для різних галузей, а також дослідження різних ідей для поліпшення ефективності моделі.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Шингалов Д.А., Мелешко Е.В., Минайленко Р.Н., Резниченко В.А. Методи автоматичного аналізу тональності контенту у соціальних мережах для виявлення інформаційно-психологічних впливів. Центрально-український науковий вісник. 2017. № 30. С. 196–202
2. Сентимент аналіз засобами нейронної мережі / К. Ялова та ін. *Математичне моделювання*. 2021. Т. 1, № 33.
3. Рудзевич А.-М. П. Методи машинного навчання в сентимент аналізі текстової інформації. URL: https://ela.kpi.ua/handle/123456789/35699 (дата звернення: 01.05.2023).
4. Анализ тональности текстов с помощью сверточных нейронных сетей. URL: https://habr.com/ru/companies/vk/articles/417767 (дата звернення: 01.05.2023).
5. Доценко С. І. Огляд методів обробки та аналізу текстів на природних мовах. Інформаційно–керуючі системи на залізничному транспорті. 2018. № 6.
6. Alex Graves, Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures// Neural Networks, 2005. – Vol. 18. – Iss.5–6.
7. Сентимент аналіз засобами нейронної мережі / К. Ялова та ін. Математичне моделювання. 2021. Т. 1, № 44.

RESEARCH AND DEVELOPMENT OF A NEURAL NETWORK MODEL FOR TEXT SENTIMENT ANALYSIS

Sergiy Platonenko

PhD, Associate Professor of IS department, Anatolii Nikolenko

Odessa Polytechnic National University, UKRAINE

ANNOTATION. A neural network model based on a recurrent neural network has been investigated and developed for text sentiment analysis.