УДК 004.932

В.А. Самойленко
Т.В. Ковалюк

МАТЕМАТИЧНА МОДЕЛЬ ТА АЛГОРИТМ РОЗПІЗНАВАННЯ ІНФОРМАЦІЙНИХ ФЕЙКІВ ЗА ДОПОМОГОЮ НЕЙРОННИХ МЕРЕЖ

Київський національний університет імені Тараса Шевченка

Анотація

У роботі запропонований метод уваго-орієнтованої двонаправленої довгої короткочасної пам'яті на основі комбінації нейронної мережі LSTM та механізму уваги для розпізнавання та класифікації фейкових новин. Запропонований метод забезпечив на 10% вищу середню точність розробленої моделі порівняно з використанням стандартної моделі нейронної мережі з довгою короткочасною пам'яттю.

Ключові слова: розпізнавання фейків, нейронні мережі, довга короткочасна пам’ять, механізм уваги

Abstract

The paper proposes a method of attention-oriented bidirectional long-term memory based on a combination of the LSTM neural network and the attention mechanism for recognizing and classifying fake news. The proposed method provided a 10% higher average accuracy of the developed model compared to the use of a standard neural network model with long short-term memory.

Keywords: fake recognition, neural networks, long short-term memory, attention mechanism

Вступ

Інформаційні фейки – це навмисно сфабриковані або суттєво спотворені повідомлення, що мають за мету ввести в оману аудиторію. Зростаюча проблема поширення фальшивих новин через соціальні мережі стала причиною занепокоєння, адже фейки можуть викликати паніку та мати серйозні наслідки, включно з фінансовими втратами. Вплив фейкових новин на війну в Україні є значним, а неправдиві відомості сприяють поширенню дезінформації, загостренню напруженості та подальшій дестабілізації в країні. Дослідження, що спрямовані на виявлення фейків, стикається з певними проблемами, зокрема складністю збирання та визначення тексту, що відноситься до фейкової новини; неефективністю блокування таких новин в месенджерах і можливості поширення фейкової інформації. З цієї причини виявлення фейкових новин стає критично важливим завданням.

Постановка задачі

Метою дослідження є покращення ефективності розпізнавання та ідентифікації інформаційних фейків в мережі Інтернет за допомогою нейромережевих засобів та технологій.

Об'єктом дослідження є новинні тексти в соціальних мережах як середовище фейків.

Предметом дослідження є моделі та алгоритми нейронних мереж, що використовуються для виявлення фейкових новин у онлайн соціальних мережах та інших мережевих ресурсах.

Завдання дослідження для досягнення мети, включають аналіз існуючих технологічних рішень і підходів до виявлення фейкових новин в мережі Інтернет; розробку методу та алгоритму виявлення та ідентифікації фейків за допомогою нейромереж; підготовка набору даних та препроцесорне оброблення для тренування нейронної мережі; проведення тренування нейронної мережі для розпізнавання фейкових новин; створення інформаційної технології для застосування розробленого методу; оцінка отриманих результатів застосування розробленого методу на основі статистичних показників.

Математична модель задачі виявлення і розпізнавання фейків

На сьогодні існують різноманітні методи та алгоритми, які застосовуються для розпізнавання фейків, зокрема такі:

- алгоритми обробки природної мови (NLP), лексичні та семантичні моделі тексту;
- алгоритми кластеризації та класифікації текстів;

- нейромережеві алгоритми, використовують різні типи нейронних мереж для виявлення фейків (LSTM, CNN, DSN, RNN, GRNN тощо)

Для виявлення фейків автори вибрали модель рекурентної нейронної мережі (RNN) з довгою короткочасною пам'яттю (LSTM). Рекурентна нейронна мережа — це тип штучної нейронної мережі, яка використовує послідовні дані або дані часових рядів. Ці алгоритми глибокого навчання зазвичай використовуються для таких задач, як переклад мови, обробка природної мови (NLP), розпізнавання мовлення та підписів до зображень тощо. Рекурентні нейронні мережі використовують навчальні дані для навчання. Вони відрізняються своєю «пам'яттю», оскільки беруть інформацію з попередніх вхідних даних, щоб впливати на поточні вхідні та вихідні дані. Таким чином, контекст даних (попередні вхідні дані) зберігається під час навчання мережі.

RNN, як правило, стикаються з двома проблемами, відомими як вибухові градієнти та зникнення градієнтів. Градієнт є нахилом функції втрат уздовж кривої помилок. Коли градієнт занадто малий, він продовжує зменшуватися, оновлюючи вагові параметри, поки вони не стануть незначними. У випадку, коли вагові параметри досягнуть 0, алгоритм більше не навчається. Вибухові градієнти виникають, коли градієнт занадто великий, що створює нестабільну модель. У цьому випадку ваги моделі виростуть занадто великими, і в кінцевому підсумку вони будуть представлені як NaN. Одним із рішень цих проблем є зменшення кількості прихованих шарів у нейронній мережі, усунення частини складності в моделі RNN.

Модель RNN з довгою короткочасною пам'яттю (LSTM) є рішенням проблеми зникнення градієнта. Суть рішення: якщо попередній стан, який впливає на поточний прогноз, не належить до недавнього минулого, модель RNN може бути не в змозі точно передбачити поточний стан. Щоб виправити це, LSTM мають «клітини» в прихованих шарах нейронної мережі, які мають три шлюзи: вхідний, вихідний і забутий. Ці шлюзи контролюють потік інформації, який необхідний для прогнозування виходу в мережі і дозволяють LSTM розглядати набагато довші вхідні послідовності. Вхідні шлюзи використовують сигмоїдну функцію, яка визначає, які значення передавати через рекурентну мережу. Нуль відкидає значення, тоді як 1 зберігає його. Після врахування поточних вхідних даних і стану пам'яті вихідний шлюз вирішує, які значення перенести на наступний часовий крок. У вихідному шлюзі визначається важливість значень в діапазоні від -1 до 1. Шлюз забуття видаляє дані, які модель LSTM вважає непотрібними для прийняття рішення щодо природи вхідних значень: 0 (забути) до 1 (зберегти). Структура LSTM подана на рис.1 [1].

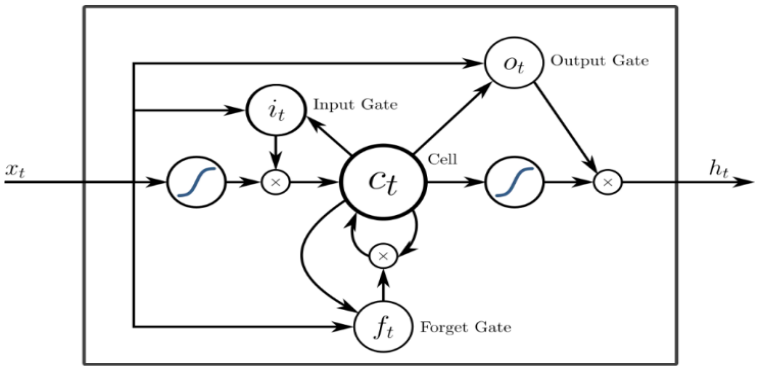


Рисунок 1 – Структура нейронної мережі LSTM

Розглянемо математичне подання моделі LSTM для класифікації фейкових новин.

Шлюз забуття f_t поданий формулою (1):

$$f_t = \sigma(W_t \cdot [h_{t-1}, x_t] + b_f), \quad (1)$$

де σ – сигмоїдальна функція активації, W_t – ваги затвора шлюза, b_f – зсув шлюза забуття, h_{t-1} – попередній вихідний сигнал, x_t – поточний вхід.

Вхідний шлюз (i_t) можна описати формулою (2):

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i). \quad (2)$$

Кандидат стану клітини ($\tilde{C}_t$) описаний формулою (3):

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c), \quad (3)$$

де W_i, W_c – ваги вхідного шлюзу і кандидата стану клітини відповідно, b_i, b_c – їх зсуви.

Оновлення стану клітини (C_t):

$$\mathrm{C}_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t, \quad (4)$$

де C_t – новий стан клітини, C_{t-1} – попередній стан клітини.

Вихідний шлюз (o_t) і кінцевий вихід (h_t) подані формулами (5) і (6):

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o \quad (5)$$

$$h_t = o_t \cdot \tanh(\mathrm{C}_t), \quad (6)$$

де W_o – ваги вихідного шлюзу, b_o – зсув вихідного шлюзу.

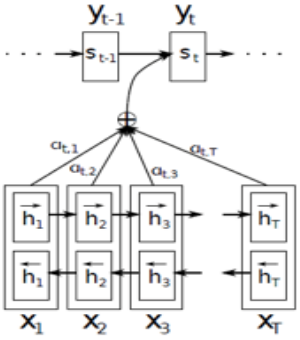
Для побудови класифікатора на основі LSTM-моделі розглянемо такі вирази. Вектор (h_t), який є вихідним станом LSTM на кожному кроці, може використовуватися для визначення, чи є введені дані, тобто новинний текст фейком. Збір вихідних станів можна агрегувати (наприклад, за допомогою усереднення або використання іншого механізму уваги) для отримання єдиного представлення новини. Кінцевий класифікаційний шар (зазвичай повнозв'язний) використовується для визначення імовірності того, що новина є фейковою, згідно з формулою (7):

$$P(\text{Фейк} \mid h) = \sigma(W_p \cdot h + b_p), \quad (7)$$

де W_p і b_p – параметри класифікаційного шару.

Ця LSTM-модель може бути навчена за допомогою набору даних із позначками «фейкова» або «справжня» для кожної новини, оптимізуючи параметри мережі таким чином, щоб мінімізувати втрати. Процедури навчання та тестування мережі LSTM складаються з трьох базових етапів: пряме розповсюдження сигналів по мережі через кожен шар, навчання мережі за допомогою алгоритму зворотного розповсюдження помилки, збереження значень сигналу пам'яті в кожному прихованому шарі як ортогональний набір у багатовимірному просторі.

Застосування механізму уваги для аналізу текстів за допомогою нейронних мереж може зважувати релевантність будь-якої області вхідного тексту та враховувати ці ваги під час реалізації запиту до тексту. Механізм уваги в загальному випадку використовує три елементи, а саме запити Q, ключі K, значення V. Запит складається із вихідних даних s_{t-1}, значення є вхідними параметрами h_i. Діаграма механізму уваги подана на рис.1 [2].



.

Рисунок 1 – Діаграма роботи механізму уваги

Перелік дій механізму уваги такий. Кожен вектор запиту $q = s_{t-1}$ зіставляється з базою даних ключів для обчислення вагової оцінки. Ця операція зіставлення обчислюється як скалярний добуток запиту, що розглядається, з кожним вектором-ключем k_i: $e_{qk_i} = q \times k_i$. Для оцінок застосовується операція *softmax* для створення вагових коефіцієнтів: $a_{qk_i} = softmax(e_{qk_i})$. Загальна увага обчислюється за допомогою зваженої суми векторів значень v_{k_i}, де кожен вектор значень поєднується з відповідним ключем: $attention(Q,K,V) = \sum_i a_{q_k i} \times v_{k_i}$. Отже, функція уваги є функцією перетворення запиту і набору пар ключ-значення на вихідну послідовність.

Опис алгоритму виявлення і розпізнавання фейків

Eman 1. Збір та попередня обробка даних

На першому етапі здійснюється збір даних з різних джерел соціальних медіа. Дані підлягають попередній обробці, яка включає нормалізацію тексту, видалення зайвих символів і стоп-слів, а також токенізацію тексту. Це дозволяє підготувати текст до подальшого аналізу та забезпечити його однорідність.

Eman 2. Векторизація тексту

Після попередньої обробки тексту виконується його векторизація, яка здійснюється за допомогою вбудованих словників. Векторизація перетворює текст в числові вектори, які можуть бути оброблені машинними алгоритмами.

Eman 3. Класифікація за допомогою двонаправленої LSTM

З векторизованих даних формується вхідний потік для моделі двонаправленої LSTM (BiLSTM). яка. обробляє послідовності векторів, аналізуючи інформацію як в прямому, так і в зворотному напрямку, що дозволяє визначити контекстуальні зв'язки в текстових даних з більшою точністю. Ця здатність забезпечує глибше розуміння смислових відносин між елементами тексту.

Eman 4. Додавання шару уваги

Після обробки тексту двонаправленою LSTM в архітектуру моделі вводиться шар уваги (attention layer). Він є додатковим шаром, що використовується в рекурентних нейронних мережах для «звернення уваги» наступних шарів мережі на прихований стан нейронної мережі в момент часу. Цей шар дозволяє моделі зосередитися на найбільш значущих словах або фразах у тексті, що важливо для точного визначення фейкових новин. Використання механізму уваги забезпечує вищу точність моделі, акцентуючи увагу на ключових елементах контенту.

Eman 5. Пост-обробка та візуалізація результатів

На останньому етапі результати класифікації проходять через процес пост-обробки, де може виконуватися додаткова перевірка та корекція результатів. Результати візуалізуються за допомогою графіків та таблиць для зручності аналізу та презентації.

Eman 6. Оцінка ефективності

Оцінка ефективності алгоритму здійснюється на основі метрик, таких як точність (accuracy), F1-оцінка, та площа під ROC-кривою (AUC). Ці метрики дозволяють оцінити здатність моделі ефективно відрізняти фейкові новини від правдивих.

Метрики оцінювання результатів та їх аналіз

1. F1-score (F1-оцінка).

F1-оцінка є гармонічним середнім значенням точності (precision) і повноти (recall). Ця метрика забезпечує одне число, що балансує компроміс між точністю і повнотою. Розраховується згідно з формулою (8):

$$F1Score = \frac{2*Precision*Recall}{\text{Precision+Recall}} \quad (8)$$

2. Accuracy (Точність).

Точність – це відношення кількості правильних прогнозів до загальної кількості зроблених прогнозів. Вона розраховується за формулою (9):

$$Accuracy = \frac{\text{True Positives+True Negatives}}{\text{Total Predictions}} \quad (9)$$

Хоча точність є поширеною метрикою, вона може ввести в оману у контексті класифікації фейкових новин, якщо набір даних є незбалансованим.

3. Recall (Повнота).

Повнота, також відома як чутливість або істинно позитивний рівень, є відношенням числа правильно класифікованих фейкових новин до загальної кількості фактичних позитивних випадків. Розраховується за формулою (10):

$$Recall = \frac{\text{True Positives}}{\text{True Positives+False Negatives}} \quad (10)$$

Повнота є особливо важливою, оскільки важливо ідентифікувати якомога більше фейкових новин для запобігання їх поширенню.

4. Precision (Точність).

Точність, також відома як позитивний прогностичний показник, є відношенням числа правильно класифікованих фейкових новин до загальної кількості позитивних прогнозів, зроблених моделлю. Розраховується згідно з формулою (11):

$$Precision = \frac{\text{True Positives}}{\text{True Positives+False Positives}} \quad (11)$$

Точність важлива, оскільки відображає здатність моделі правильно ідентифікувати фейкові новини, не класифікуючи помилково справжні новини як фейкові.

5. Loss (Втрати).

Втрати вимірюють, наскільки добре прогнози моделі відповідають цільовим значенням. Зазвичай для задач бінарної класифікації, як класифікація фейкових новин, використовується втрата бінарної крос-ентропії. Розраховується так:

$$f(x) = -\frac{1}{N} * \sum_{i=1}^{N} [y_i * \log(\widehat{y_i}) + (1 - y_i) * \log(1 - \widehat{y_i})], \quad (12)$$

де N – кількість зразків, y_i – фактичне цільове значення і-го зразка, $\widehat{y_i}$ – прогнозоване значення для і-го зразка.

Аналіз результатів контрольного прикладу

Експериментальні дослідження проведені на датасеті WELFake, який складається з 72134 новинних статей, з яких 35028 є справжніми, а 37106 – є фейковими.

Ключові характеристики датасету: статті, що містять від 450 до 550 слів, зазвичай є більш надійними; коротші, але змістовні статті часто виявляються більш правдивими; фейкові новини мають гіршу читабельність порівняно з справжніми новинами; фейкові статті мають більшу суб'єктивність порівняно з справжніми статтями; кількість статей, що представляють справжні новини, перевищує кількість фейкових.

Таблиця 1 ілюструє отримані результати навчання моделей LSTM та запропонованого методу для класифікації фейкових новин за допомогою ключових метрик оцінки ефективності.

Таблиця 1 – Аналіз результатів тестування

Метрика	LSTM	Запропонований метод
F1-Score	0.87431	0.97689
Accuracy	87.45%	97.72%
Recall	88.52%	97.75%
Precision	86.78%	97.82%
Loss	0.17521	0.07289

Таблиця 2 – Аналіз результатів навчання

Метрика	LSTM	Запропонований метод
F1-Score	0.89891	0.99198
Accuracy	89.97%	99.25%
Recall	90.01%	99.30%
Precision	89.91%	99.28%
Loss	0.13088	0.02499

За даними, що подані в таблицях результатів, можна зробити такі узагальнення.

Загальна ефективність: обидві моделі показали високу ефективність на навчальних та тестових наборах даних. Запропонований метод загалом продемонструвала перевагу в більшості метрик, особливо на навчальних даних.

Тестові та навчальні результати: різниця в показниках продуктивності між двома моделями на тестових даних була незначною, що свідчить про те, що обидві моделі є надійними та ефективними для класифікації фейкових новин.

Висновки

В рамках дослідження було запропоновано метод виявлення фейкових новин, який інтегрує увагоорієнтовану двонаправлену довготривалу короткочасну пам'ять в моделях, основаних на LSTM нейронних мережах. Цей метод використовує переваги механізму уваги для зосередження на важливих

сегментах тексту, що забезпечує глибше розуміння контексту та підвищує точність класифікації. Результати досліджень підтверджують значущість використання механізмів уваги в моделях глибокого навчання для завдань класифікації, особливо у контексті виявлення фейкових новин, де важливо розуміти та аналізувати контекст та вагомість окремих частин тексту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Nelson D. What are RNNs and LSTMs in Deep Learning? [Електронний ресурс]. – Режим доступу: https://www.unite.ai/uk/what-are-rnns-and-lstms-in-deep-learning/

2. A Comprehensive Guide to Attention Mechanism in Deep Learning for Everyone. [Електронний ресурс]. – Режим доступу: https://www.analyticsvidhya.com/blog/2019/11/comprehensive-guide-attention-mechanism-deep-learning/

3. Антіпова К. О. Застосування механізму уваги типу multi-head та моделі трансформера для задачі машинного перекладу. ВІСНИК ХНТУ № 1(84), 2023, с. 118 – 122.

Самойленко Владислав Андрійович, група ІПЗ-21м, факультет інформаційних технологій, Київський національний університеті імені Тараса Шевченка, м. Київ, e-mail: kekonmonday@knu.ua

Ковалюк Тетяна Володимирівна, к.т.н., доцент, доцент кафедри програмних систем і технологій, Київський національний університеті імені Тараса Шевченка, м. Київ, e-mail: tetyana.kovalyuk@gmail.com

Vladyslav Samoilenko, group IPZ-21m, faculty of information technologies, Taras Shevchenko National University of Kyiv, Kyiv, e-mail: kekonmonday@knu.ua

Tetiana Kovaliuk, Ph.D., Associate Professor, Associate Professor of the Department of Software Systems and Technologies, Taras Shevchenko National University of Kyiv, Kyiv, e-mail: tetyana.kovalyuk@gmail.com