DOI 10.34185/1562-9945-4-147-2023-14
УДК 004.9

К.Ю. Островська, І.В. Стовпченко, Д.С. Печений

ДОСЛІДЖЕННЯ МЕТОДІВ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ ДЛЯ АНАЛІЗУ ТОНАЛЬНОСТІ КОРПУСУ ТЕКСТІВ

Анотація Об'єктом дослідження є методи з урахуванням нейронних мереж для аналізу тональності корпусу текстів. Для досягнення поставленої в роботі мети необхідно вирішити такі завдання: дивчити теоретичний матеріал для навчання глибинних нейронних мереж та їх особливості стосовно обробки природної мови; вивчити документацію бібліотеки Tensorflow; розробити моделі згорткової та рекурентної нейронних мереж; розробити реалізацію лінійних та нелінійних методів класифікації на моделях мішка слів та Word2Vec; порівняти точність та інші показники якості реалізованих нейромережевих моделей із класичними методами. Для візуалізації навчання використовується Tensorboard. У роботі показано перевагу класифікаторів на основі глибоких нейронних мереж над класичними методами класифікації, навіть якщо для векторних уявлень слів використовується модель Word2Vec. Найвищу точність для даного корпусу текстів має модель рекурентної нейронної мережі з LSTM блоками.
Ключові слова: штучні нейронні мережі, ГЛибокі нейронні. мережі, навчання з учителем, глибоке навчання, рекурентна нейронна мережа, LSTM, згорткова нейронна мережа, аналіз тональності тексту, мішок слів, Word2vec.

Вступ. Щодня в Інтернеті з'являється величезна кількість контенту: користувачі висловлюють свою думку про фільми та події, залишають відгуки про різні продукти та послуги. Для вирішення завдань, пов'язаних із аналізом емоційного забарвлення тексту, використовуються методи аналізу тональності тексту.

Завдання автоматичного аналізу тональності тексту є досить популярним [1]. Найчастіше користувачі при виборі чогось (наприклад, до якого університету вступити) керуються думками інших людей. Тому набір опрацьованих думок становить значний інтерес для соціологів, маркетологів та власників бізнесу. Також система для аналізу тональності тексту може стати в нагоді на сайтах, де люди залишають відгуки: наприклад, користувач може позитивно оцінити фільм, випадково поставивши негативну оцінку - система автоматичного аналізу тональності тексту виправить помилку. Тому аналіз тональності текстів є важливим та актуальним завданням.

Автоматичний аналіз тональності тексту зазвичай застосовується на корпусах текстів, які містять рецензії чи відгуки. Також його можна застосувати для даних з блогів та соціальних мереж, щоб отримати громадську думку з того чи іншого питання або товару.

Визначення емоційної оцінки авторів рецензій перестав бути очевидним завданням. Стандартні методи класифікації тексту ізолюють слова за ознаками та застосовують різні методи вибору ознак, щоб знайти найбільш “важливе слово” у тексті: наприклад, речення "Мені подобається бігати" і "Мені не подобається бігати" вважатимуться однаковими. У ситуації, коли рецензія представлена лише позитивно забарвленим набором слів, а насправді негативною, стандартні методи класифікації тексту перестають бути ефективними. Методи машинного навчання дозволяють алгоритмам “розуміти” структуру речення та його семантичну структуру. Пропозиція буде представлена у вигляді вектора, в якому збережена структура речення та те, як слова пов'язані один з одним.

Для проведення чисельних експериментів були використані рецензії сайту Rotten Tomatoes [2] — набір із 5331 позитивних та 5331 негативних рецензій.

Потрібно побудувати бінарний класифікатор, визначальною, позитивною чи негативною виявилася рецензія. Як методи розглядаються згорткова нейронна мережа і рекурентна нейронна мережа з LSTM-блоками, наївний класифікатор байесовського, метод опорних векторів і логістична регресія. Також було використано два варіанти векторних моделей подання тексту: мішок слів (Bag of Words) та Word2Vec.

В результаті навчання були отримані моделі нейронних мереж, що дозволяють з високою точністю визначати тональність рецензій, а також порівняння ефективності використання реалізованих методів.

Аналіз останніх досліджень і публікацій. Завданням, поставленою у роботі, є реалізація різних алгоритмів глибинного навчання для аналізу тональності тексту та порівняння їх ефективності з класичними (лінійними та нелінійними) алгоритмами класифікації тексту.

У дослідженні [3] представлені результати ефективності застосування різних архітектур згорткових нейронних мереж при використанні попередньо навченої моделі векторного уявлення word2vec.

Результати застосування нейронних мереж з LSTM-блоками та їх модифікаціями були представлені в дослідженні [4], а також було показано, що ця архітектура є найбільш ефективною для побудови бінарного та багатокласового класифікаторів для аналізу тональності текстів.

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

Також у роботі [5] стверджується, що можна значно покращити ефективність класифікатора, використовуючи попередньо навчені моделі векторних уявлень слів (наприклад, Word2Vec).

У цій роботі для побудови бінарного класифікатора будуть використовуватися рецензії, що складаються з коротких речень.

В даний час завдання побудови класифікатора (бінарного або багатокласового) для визначення тональності тексту вирішується за допомогою нейромережевих моделей, оскільки ефективність архітектур, використаних у згаданих роботах, значно вища, ніж у класичних лінійних алгоритмів.

Існує безліч бібліотек для реалізації алгоритмів машинного навчання. Існують два типи фреймворків: символьні та імперативні. У символьних фреймворках набагато більше можливостей використовувати пам'ять багаторазово, а оптимізація на основі граф залежностей здійснюється автоматично. Найпопулярнішими символьними (symbolic) фреймворками нині є TensorFlow та Theano.

На відміну від Theano, Tensorflow не орієнтований лише на навчання нейронних мереж, тому можна використовувати колекції графів та черги як складові для високорівневих компонентів.

Якщо необхідно навчати масштабні моделі і використовувати багато зовнішньої пам'яті, Theano буде дуже повільно працювати через необхідність компіляції коду C/CUDA в бінарний код.

TensorFlow має прозору модульну архітектуру з безліччю фронт-ендів. В архітектурі Theano розібратися досить непросто: весь код – це Python, де код C/CUDA упакований як рядок Python. У такому коді складно орієнтуватися, його непросто налагоджувати та проводити рефакторинг. Більше того, візуалізація графів у TensorFlow реалізована значно ефективніше, ніж у Theano [6].

Векторні уявлення слів для лінійних алгоритмів будуть представлені двома моделями: Word2Vec та мішок слів (bag of words), а для класифікаторів на основі нейронних мереж буде використано лише модель Word2Vec. За допомогою інструменту для побудови векторних моделей Gensim буде навчено модель Word2Vec. З використанням TensorFlow будуть реалізовані згорткова нейронна мережа та рекурентна нейронна мережа з LSTM-блоками.

Методи. Процес обробки даних складається з наступних кроків:

- видалити розмітку HTML;
- видалити всі символи, окрім літер та пробілів;

- з отриманого набору слів видалити стоп-слова.

Наступним завданням є перетворення кожної рецензії на векторне уявлення. Для оцінки ефективності кожного з методів буде використано дві моделі векторного подання слів: мішок слів та Word2Vec.

У роботі логістична регресія використовується для передбачення ймовірності виникнення певної події.

Наївний класифікатор Байєса - простий імовірнісний класифікатор, заснований на застосуванні Теореми Байєса зі суворими припущеннями про незалежність елементів вектора ознак. Перевагою наївного класифікатора Байєса є мала кількість даних для навчання, необхідних для оцінки параметрів, необхідних для класифікації.

Випадковий ліс (random forest) – алгоритм полягає у використанні комітету вирішальних дерев. Класифікація об'єктів проводиться шляхом голосування: кожне дерево комітету відносить об'єкт, що класифікується, до одного з класів, і перемагає клас, за який проголосувало найбільшу кількість дерев. Оптимальна кількість дерев підбирається таким чином, щоб мінімізувати помилку класифікатора на тестовій вибірці. Використовує усереднення для підвищення точності прогнозування та контролю надлишкового припасування. Розмір підвиборки завжди збігається з розміром оригінальної вибірки.

Метод опорних векторів (SVM) - Основна ідея методу - переведення вихідних векторів у простір більш високої розмірності та пошук роздільної гіперплощини з максимальним зазором у цьому просторі. Стандартна функція scikitlearn (sklearn.svm.SVC) не підходить, оскільки тимчасова складність даного алгоритму квадратична. Ефективність методу опорних векторів значно знижується, якщо кількість ознак дуже велика. Він має велику гнучкість у виборі штрафів та функцій втрат і має краще масштабуватись для великої кількості зразків. Оптимізація функції втрат SVM за допомогою градієнтного спуску:

$$L(\omega, D) = \frac{1}{2}\|\omega\|_2^2 + C\sum_{i=1}^{N} max(0, 1 - y_i(\omega^T x_i + b)), \quad (1)$$

де

$$D = \{(x_i, y_i)\}_{i=1}^{N}, x_i \in Rd \text{ and } y_i \in \{-1, +1\}. \quad (2)$$

Оптимізація функції втрат - (regularization_loss + hinge_loss). Для випадку word2vec буде використано SGDClassifier зі sklearn.linear_model з помилкою l2.

Згорткова нейронна мережа - Векторні подання даних здійснено за допомогою моделі Word2Vec (Skip-gram Model).

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

Як функція оптимізації використовується Adam (Adaptive Moment Estimation). Його відмінними рисами є:

- оцінка першого моменту обчислюється як ковзне середнє;

- оскільки оцінки першого та другого моментів ініціалізуються нулями, використовується невелика корекція, щоб результуючі оцінки не були зміщені до нуля.

Рекурентна нейронна мережа з LSTM-блоками - векторні подання даних здійснено за допомогою моделі Word2Vec (Skip-gram Model). LSTM вважаються найкращою архітектурою для аналізу тональності тексту. Нейронні мережі, складені з LSTM-модулів, особливо добре обробляють заперечення (negation), якщо у осередку є projection unit (тобто більше пам'яті мережі).

Як досвідчений зразок були використані рецензії веб-сайту RottenTomatoes [1] — набір із 5331 позитивних та 5331 негативних рецензій. Для розробки використовувалися бібліотеки Pandas, Scikit-Learn та PyMorphy2, а також фреймворк TensorFlow як засіб аналізу.

Алгоритм аналізу даних у додатку.

1. Отримуємо дані з іншого джерела (бази даних, користувальницького інтерфейсу).

2. Видаляємо зайву інформацію з пропонованого тексту залишаючи тільки російські літери.

3. Проводимо морфологічний аналіз тексту, та лематизуємо текст.

4. Будуємо модель:

- схема n-грам: (1, 3) (уніграми + біграми + триграми);

- Метод векторизації: Word2Vec;

- Тип моделі: Рекурентна нейронна мережа з LSTM-блоками;

Параметри моделі: penalty – l2, alpha – 0.000001, loss – log.

5.Навчаємо нейронну мережу за отриманими даними.

Алгоритм навчання нейронної мережі. Для того, щоб нейронна мережа почала виконувати свої завдання її необхідно навчити, процес навчання відбувається за принципом, показаним на рисунку 1.

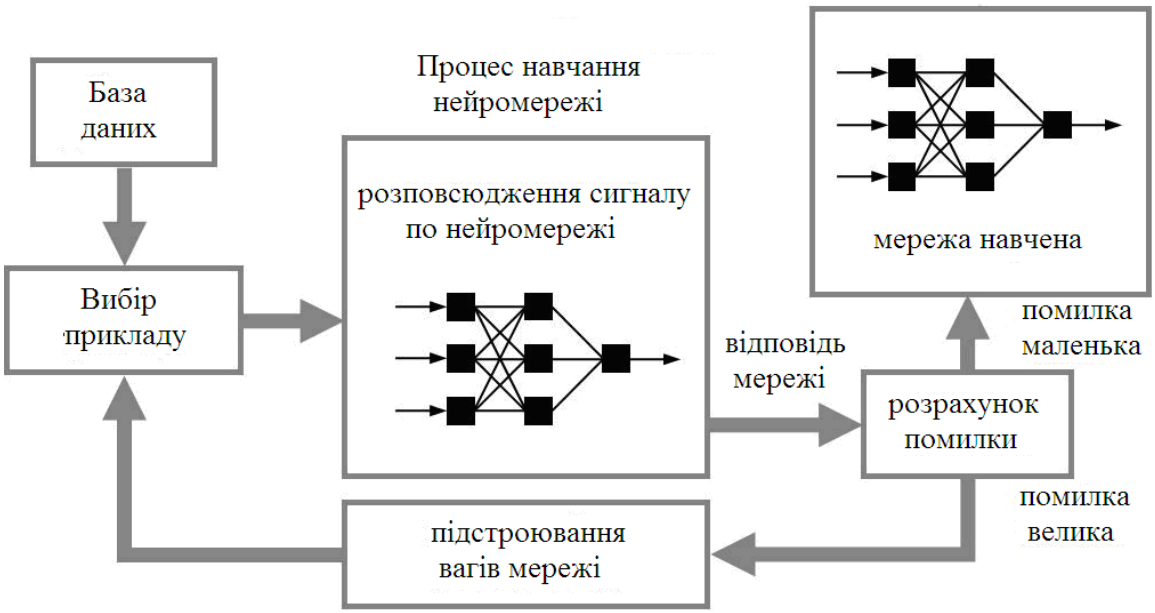

Рисунок 1 - Частка коректних прогнозів (accuracy):
синій графік - навчання, червоний – перевірка

Показники якості нейронної мережі. Частка коректних прогнозів (accuracy) - відсоток помилок, допустимих класифікатором.

Наступні показники будуть використані лише для класичних методів класифікації:

• міра точності (precision) - відношення tp до (tp + fp), де tp - кількість істинних позитивних величин, а fp – кількість хибних позитивних величин. Тобто міра точності характеризує скільки одержаних від класифікатора позитивних рішень вважаються вірними;

• Міра повноти (recall) – відношення tp до (tp + fn), де fn – кількість хибних негативних величин. Міра повноти встановлює вміння класифікатора дізнаватися так само як і найбільше позитивних рішень з прогнозованих;

• Міра F1 – середній гармонійний міри точності та міри повноти.

Визначає лімінальну властивість класифікатора;

• Носій міри (support) – кількість інформації будь-якого з класів.

Найбільш жорстке визначення: мінімальне закрите велике число, в якому сконцентровано ступінь.

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

Метод мішка слів (див. таблиці 1 - 4).

Таблиця 1

Логістична регресія

Клас	Мера точності	Мера повноти	Мера F1	Носій міри
0	0.75	0.76	0.75	1192
1	0.74	0.74	0.74	1139
total	0.74	0.74	0.74	2331
Достовірність 0.756				

Таблиця 2

Наївний байєсовський класифікатор

Клас	Мера точності	Мера повноти	Мера F1	Носій міри
0	0,73	0,73	0,72	1192
1	0,71	0,75	0,72	1139
total	0,72	0,71	0,71	2331
Достовірність 0.730				

Таблиця 3

Випадковий ліс

Клас	Мера точності	Мера повноти	Мера F1	Носій міри
0	0,71	0,74	0,72	1192
1	0,71	0,68	0,69	1139
total	0,71	0,71	0,71	2331
Достовірність 0.718				

Таблиця 4

Лінійний метод опорних векторів

Клас	Мера точності	Мера повноти	Мера F1	Носій міри
0	0,79	0,56	0,65	1192
1	0,64	0,85	0,73	1139
total	0,72	0,70	0,69	2331
Достовірність 0.689				

Метод Word2Vec (див. таблиці 5 - 8).

Таблиця 5

Логістична регресія

Клас	Мера точності	Мера повноти	Мера F1	Носій міри
0	0.77	0.77	0.77	1192
1	0.76	0.76	0.76	1139
total	0.86	0.86	0.86	2331
Достовірність 0.767				

Таблиця 6

Наївний байєсовський класифікатор

Клас	Мера точності	Мера повноти	Мера F1	Носій міри
0	0.73	0.72	0.72	1192
1	0.71	0.72	0.71	1139
total	0.72	0.72	0.72	2331
Достовірність 0.728				

Таблиця 7

Випадковий ліс

Клас	Мера точності	Мера повноти	Мера F1	Носій міри
0	0.73	0.74	0.75	1192
1	0.73	0.73	0.73	1139
total	0.73	0.73	0.74	2331
Достовірність 0.738				

Таблиця 8

Лінійний метод опорних векторів

Клас	Мера точності	Мера повноти	Мера F1	Носій міри
0	0.831	0.633	0.711	1092
1	0.691	0.864	0.771	1039
total	0.762	0.745	0.741	2131
Достовірність 0.743				

Згорткова нейронна мережа. Достовірність 0.789. Графіки точності моделі та функції втрат представлені на рисунках 2 та 3 відповідно.

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

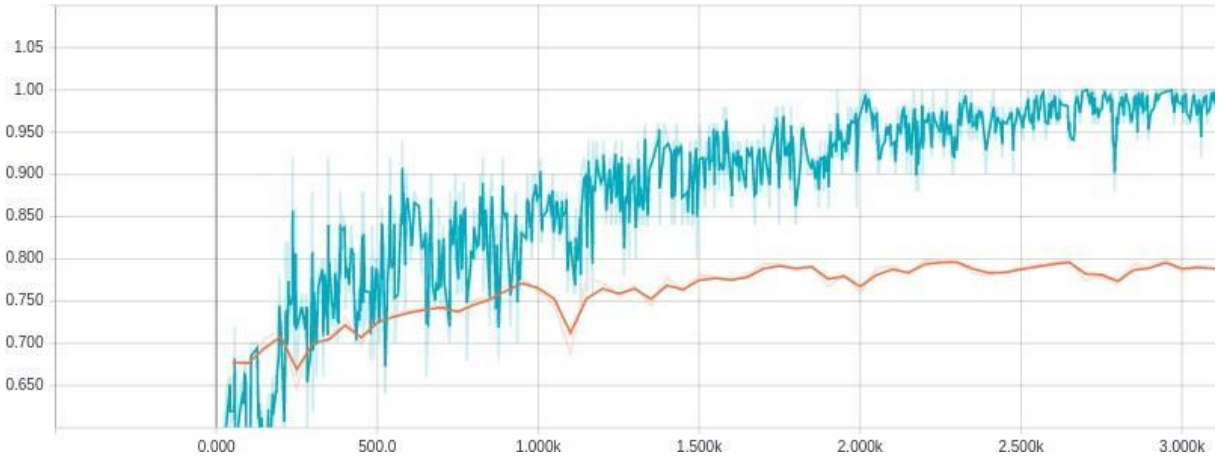

Рисунок 2 - Частка коректних прогнозів (accuracy):
синій графік - навчання, червоний – перевірка

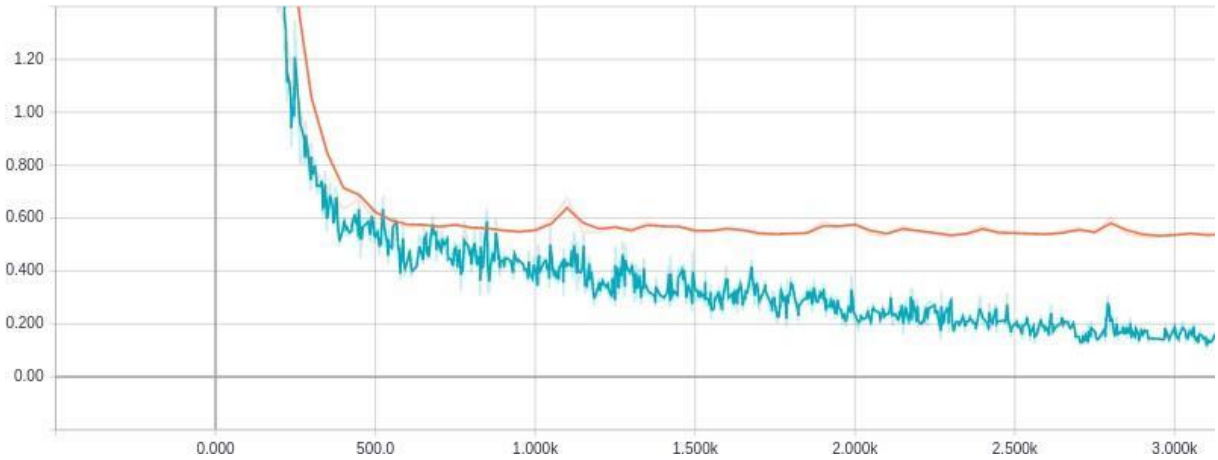

Рисунок 3 - Функція втрат: синій графік - навчання, червоний - перевірка

Рекурентна нейронна мережа з LSTM-блоками. Вирішальним у навчанні моделі виявився вибір функції мінімізації градієнтного спуску. Спочатку було використано алгоритм RMSProp (root mean square propogation), ідея якого полягає у масштабуванні градієнта.

Однак за інших рівних умов (розміру прихованого LSTM шару та темпі навчання) застосування алгоритму оптимізації Adam (який крім ідеї масштабування градієнта використовує ідею інерції) дозволило досягти максимальної точності щодо вже реалізованих методів цій роботі - 83.1%.

Графіки точності моделі та функції втрат представлені на рисунках 4 та 5 відповідно.

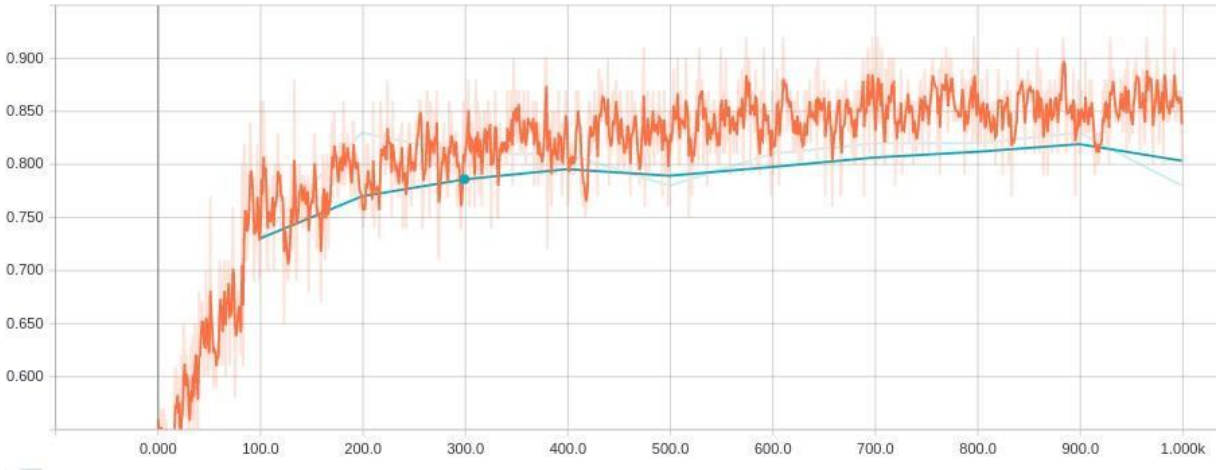

Рисунок 4 - Частка коректних прогнозів (accuracy):
червоний графік-навчання, синій – проверка

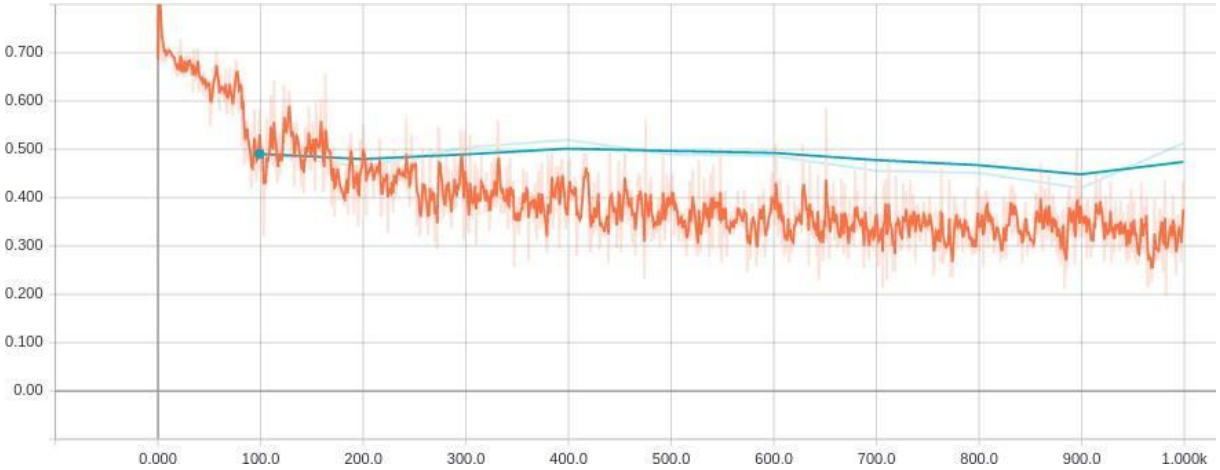

Рисунок 5 - Функція втрат: червоний графік - навчання, синій – перевірка

Результати. Для класичних алгоритмів класифікації використання мішка слів як модель векторного подання слів точність передбачень не перевищувала 75.4% (логістична регресія), мінімальна точність становила 69.9% (лінійний метод опорних векторів).

Завдяки використанню моделі Word2Vec вдалося покращити точність передбачень майже всім методів (крім наївного байєсівського класифікатора): наприклад, точність лінійного методу опорних векторів стала 74.2%. Проте найефективнішим бінарним класифікатором виявилася логістична регресія: її точність із використанням моделі Word2Vec дорівнює 76,6%. За допомогою згорткових нейронних мереж з моделлю Word2Vec вдалося отримати точність 79.9%.

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

Найефективнішою архітектурою для аналізу тональності тексту виявилася рекурентна нейронна мережа з LSTM-блоками. Її максимальна точність становила 83.0%.

Отримані експериментальні дані показують вищу ефективність роботи глибоких нейронних мереж проти класичними алгоритмами для аналізу тональності тексту.

Результати дослідження свідчать, що використання глибоких нейронних мереж значно покращує точність аналізу тональності тексту.

Перевага рекурентної мережі на основі LSTM над згортковою нейронною мережею в галузі аналізу тональності вже була доведена в різних дослідженнях, проте важливо відзначити, що в даній роботі були реалізовані найпростіші архітектури глибоких нейронних мереж. Поліпшення параметрів моделі, використання більш розширеної моделі векторного уявлення слів Word2Vec, застосування attention-механізмів дозволить значно збільшити ефективність бінарного класифікатора для аналізу тональності на основі глибоких нейронних мереж.

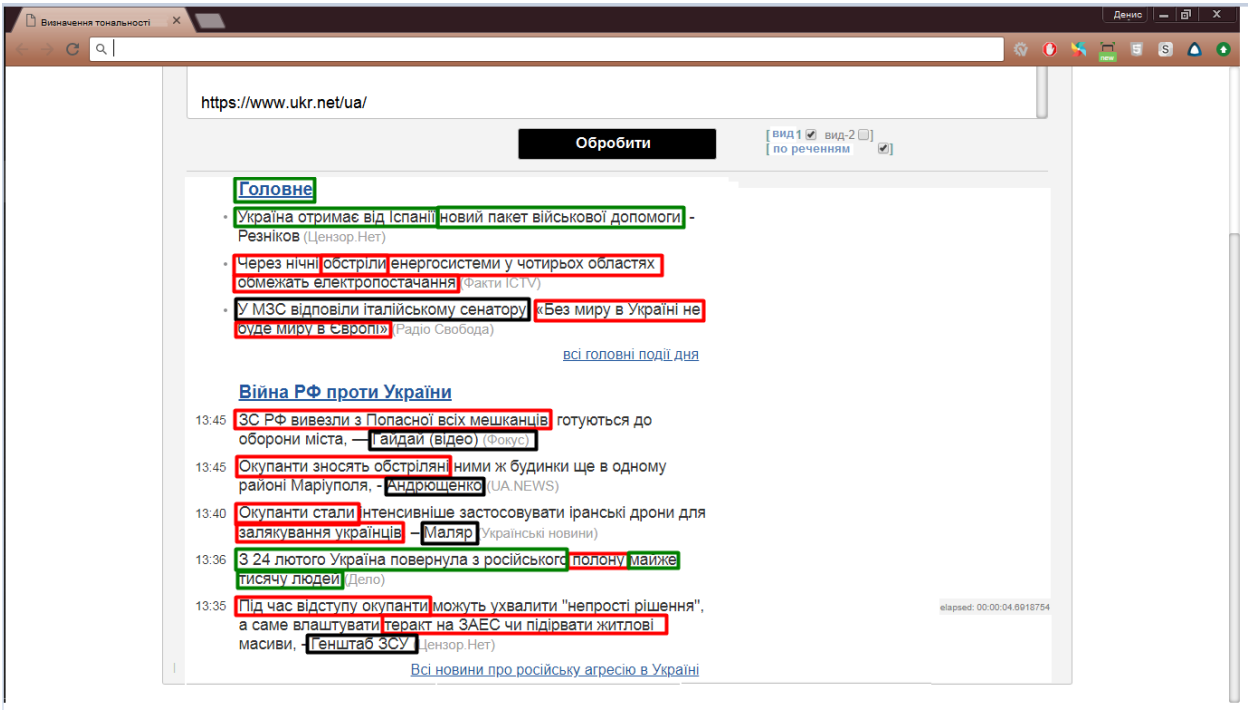

Рисунок 6 - Приклад роботи модуля

Перспективи подальших досліджень. Можливим напрямом для подальшої роботи може бути розпізнавання ключових слів, які роблять найбільший внесок у позитивний або негативний відгук. Введення модуля в експлуатацію.

Висновки. Відповідно до поставленої мети - розробка алгоритмів глибокого навчання для аналізу тональності тексту та порівняння їх ефективності з

іншими класифікаторами на основі алгоритмів машинного навчання – було реалізовано архітектуру згорткової нейронної мережі, рекурентної нейронної мережі з LSTM-блоками, а також проведено порівняння показників якості їх класифікації з іншими класифікаторами.

При використанні моделі мішка слів точність різних методів була значно вищою за випадкову (близько 70%), проте застосовуючи модель Word2Vec, вдалося значно покращити точність роботи алгоритмів (на кілька одиниць). Однак нейронні мережі показали найкращі результати.

Точність класифікатора на основі згорткової нейронної мережі виявилося 79.9%. Найвищу точність показав класифікатор на основі рекурентної мережі з LSTM-блоками – 83.3%.

ЛІТЕРАТУРА / REFERENCES

1. B. Pang and L. Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval archive, 2008.

2. Данные рецензий, используемых в работе, sentence polarity dataset v1.0. http://www.cs.cornell.edu/people/pabo/movie-review-data/

3. Y.Kim. Convolutional neural networks for sentence classification.arXiv:1408.5882 [cs.CL], 2014.

4. K.S. Tai et al. Improved semantic representations from tree-structured long short-term memory network. arXiv:1503.00075 [cs.CL], 2015.

5. Q. Le and T. Mikolov. Distributed representations of sentences anddocuments. arXiv:1405.4053 [cs.CL], 2014.

6. K.Tran et al. Evaluation of deep learning toolkits. https://github.com/zer0n/deepframeworks/blob/master/README.md

Received 19.05.2023.
Accepted 24.05.2023.

Research of methods based on neural networks for the analysis of the tonality of the corps of the texts

The object of the study is methods based on neural networks for analyzing the tonality of a corpus of texts. To achieve the goal set in the work, it is necessary to solve the following tasks: study the theoretical material for learning deep neural networks and their features in relation to natural language processing; study the documentation of the Tensorflow library; develop models of convolutional and recurrent neural networks; to develop the implementation of linear and non-linear classification methods on bag of words and Word2Vec models; to compare the accuracy and other quality indicators of implemented neural network models with classical methods. Tensorboard is used for

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

learning visualization. The work shows the superiority of classifiers based on deep neural networks over classical classification methods, even if the Word2Vec model is used for vector representations of words. The model of recurrent neural network with LSTM blocks has the highest accuracy for this corpus of texts.

Keywords: artificial neural networks, DEEP neural networks. networks, tutored learning, deep learning, recurrent neural network, LSTM, convolutional neural network, text tonality analysis, bag of words, Word2vec.

Островська Катерина Юріївна – к.т.н., доцент, доцент кафедри інформаційних технологій і систем, Український держаний університет науки та технологій Дніпро, Україна.

Стовпченко Іван Володимирович – ст. викладач кафедри інформаційних технологій і систем, Український держаний університет науки та технологій Дніпро, Україна.

Печений Денис Сергійович - магістрант кафедри інформаційних технологій і систем, Український держаний університет науки та технологій Дніпро, Україна.

Ostrovska Kateryna - Ph.D., Associate Professor, Associate Professor of the Department of Information Technologies and Systems, Ukrainian State University of Science and Technology of Dnipro, Ukraine.

Stovpchenko Ivan - senior teacher of the Department of Information Technologies and Systems, Ukrainian State University of Science and Technology, Dnipro, Ukraine.

Pechenyi Denys - master's student at the Department of Information Technologies and Systems, Ukrainian State University of Science and Technology, Dnipro, Ukraine.