УДК 681.12

А. О. Переродов
О.К. Колесницький
І.К. Денисов

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ КЛАСИФІКАЦІЇ БАНКІВСЬКИХ ТЕКСТІВ НА ОСНОВІ ЗГОРТКОВОЇ НЕЙРОННОЇ МЕРЕЖІ

[1] Вінницький національний технічний університет;

Анотація

Запропоновано інформаційну технологію класифікації банківських текстів на основі згорткової нейронної мережі, яка для перетворення слова в вектор використовує метод word2vec, що дозволило підвищити достовірність класифікації текстів.

Ключові слова: класифікація текстів, інформаційна технологія, згорткова нейронна мережа.

Abstract

The information technology of banking text classification based on convolutional neural network is proposed, which uses word2vec method for transformation of word into vector, which allowed to increase the accuracy of text classification.

Keywords: text classification, information technology, convolutional neural network.

Вступ

В ході аналізу проблеми класифікації текстів у різних сферах діяльності було встановлено, що класифікація текстових документів для їх подальшого перенаправлення у відповідні відділи в банківських установах є досить актуальною проблемою через велику кількість вхідної кореспонденції, яка призводить до перезавантаженості служб банківського моніторингу. Під вхідною кореспонденцією мається на увазі листи, звернення, скарги та інші текстові документи отримані на адресу банку. Перезавантаженість виникає через додаткове візування вхідної кореспонденції людиною, після того як вони були класифіковані програмно. Підвищення точності класифікації дозволить вирішити проблему додаткового візування текстових документів.

Метою роботи є розроблення інформаційної технології та методу класифікації банківських текстів на основі згорткової нейронної мережі, які мають підвищену достовірність класифікації текстів.

Результати дослідження

На сьогоднішній день розроблено велику кількість методів класифікації текстів і їх різних варіацій. Кожна група методів має свої переваги і недоліки, області застосування, особливості та обмеження. Основними методами класифікації текстів є:

- наївний баєсівський класифікатор;
- метод k-найближчих сусідів;
- дерева рішень;
- метод опорних векторів;
- методи на основі штучних нейронних мереж.

Було обрано як найперспективніші, методи на основі штучних нейронних мереж.

Етапи процесу класифікації текстів зображено на рисунку 1.1.

В ході аналізу методів перетворення слова у вектор фіксованої довжини було розглянуто такі методи, як one-hot encoding, word2vec, glove. Для вирішення задачі класифікації текстів для перетворення слова в вектор був обраний метод word2vec [1], так як він забезпечує схожість векторів семантично близьких слів та має відносно невелику розмірність, що зменшує складність обчислень.

У даній роботі необхідно реалізувати програму класифікації банківських текстів. Класифікувати об'єкт – значить, вказати номер (або найменування класу), до якого відноситься даний об'єкт.

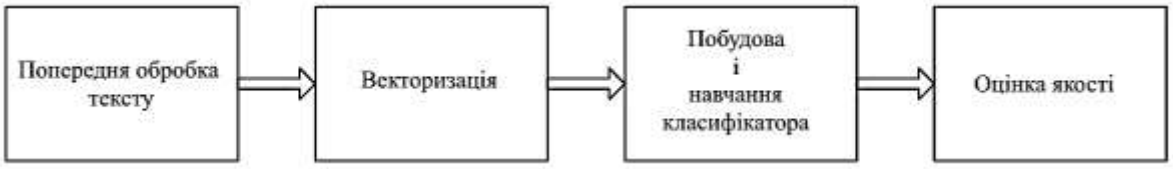

Рисунок 1.1 – Етапи процесу класифікації текстів

В задачі класифікації текстів об'єктами є текстові документи. Формальна постановка задачі класифікації текстів виглядає наступним чином:

$D = \{ d_1, \dots, d_n \}$ – множина текстових документів. Кожний документ $d \in D$ представляє собою послідовність слів $W_d = \{w_1, \dots, w_{n_d}\}, n_d$ – довжина документа d.

$Y = \{y_1, \dots, y_n\}$ – кінцева множина міток класів.

$y^*: D \to Y$ – невідома цільова залежність, значення якої відомі тільки на об'єктах кінцевої навчальної вибірки $D^m = \{(d_1, y_1), \dots, (d_m, y_m)\}$.

Потрібно розробити інформаційну технологію, в якій буде реалізовано алгоритм $a: D \to Y$, який здатний класифікувати довільний об'єкт $d \in D$.

Проаналізувавши вихідні дані, сучасні методи класифікації текстів та існуючі програми-аналоги можна сформулювати функціональні вимоги до програми класифікації банківських текстів.

Програма класифікації банківських текстів повинна виконувати класифікацію тексту на англійській мові з банківської тематики, обсягом не більше 400 слів. Програма повинна працювати на основі згорткової нейронної мережі, проводити попередню обробку тексту (видалення стоп-слів та стоп-символів) та word2vec векторизацію, що підвищать точність класифікації.

Програма класифікації банківських текстів повинна працювати як веб-сервіс (Web API) через стандартні HTTP-запити та використовувати аутентифікацію по токену. Веб-сервіс повинен підтримувати такий формат обміну даними як JSON. Веб-сервіс повинен працювати під управлінням операційної системи Windows XP/7/8/10/Server, Linux, які є найбільш поширеними серед операційних систем. Ємність ОЗУ залежить від навантаження на веб-сервер, але повинно бути не меншою за 512 Мбайт.

Надійність функціонування програми і її функціональну стійкість визначають вхідні дані, які передаються в тілі HTTP-запиту, тому необхідно передбачити перевірку вхідних даних на правильність.

Інформаційна модель процесу класифікації текстів – це модель, що описує істотні для даного процесу параметри та змінні величини, зв'язки між ними, та його вхідні та вихідні значення. Виходячи з визначення інформаційної моделі, її можна подати у вигляді кортежа:

$$IMTC = < TXT, ss, sw, va, ca, R >, \text{ де} \quad (1)$$

TXT– текст, який потрібно класифікувати,
ss – множина стоп-символів,
sw – множина стоп-слів,
va – алгоритм векторизації,
ca – алгоритм класифікації,
R – результат класифікації.

Схему інформаційної моделі процесу класифікації текстів зображено на рисунку 2.

Було визначено архітектуру згорткової нейронної мережі [2] для класифікації текстів та загальну математичну модель згорткової нейронної мережі. Мережа складатиметься з вхідного шару розмірністю 400×100, 3 згорткових шарів з фільтрами 3×100, 4×100 та 5×100, 3 агрегувальних шарів та повнозвязного вихідного шару. Кількість фільтрів для кожного згорткового шару – 32. Функція, яка буде використана в шарах агрегування – максимізаційна (max-pooling). Для розподілення ймовірності між класами для вихідного шару як функція активації буде використана функція Softmax.

Для згорткових та агрегувальних шарів функція активації – ReLU. Метод, за яким буде проходити навчання мережі – це метод зворотного поширення помилки, в основі якого лежить стохастичний градієнтний спуск.

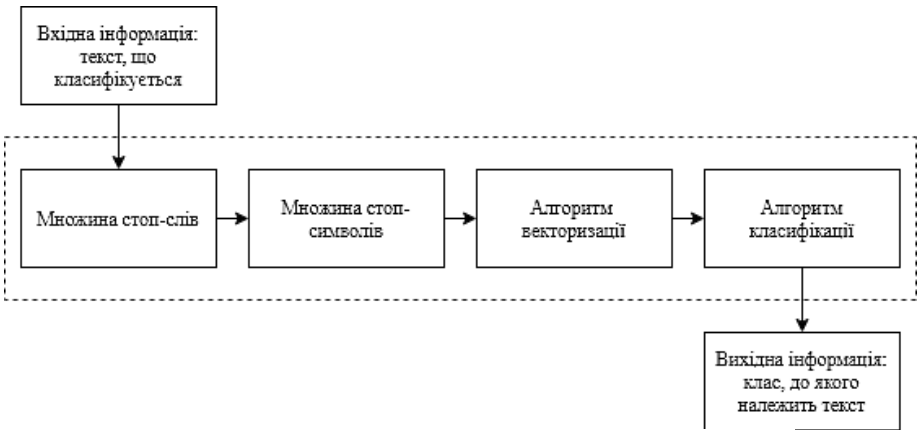

Рисунок 2 – Схема інформаційної моделі процесу класифікації текстів

В ході проектування програмного забезпечення класифікації текстів було визначено доцільність декомпозиції та її переваги у процесі проектування. На основі ряду задач, які будуть виконуватись програмним забезпеченням класифікації текстів, було проведено його декомпозицію на наступні модулі:

- модуль попередньої обробки тексту;
- модуль векторизації тексту;
- модуль класифікації;
- модуль аутентифікації.

В ході практичної реалізації інформаційної технології класифікації банківських текстів було обрано такі мови програмування: для модуля класифікації текстів – Python, для модуля аутентифікації – C#, для клієнтського веб-додатку тестування роботи WebAPI – Typescript та бібліотеки Tensorflow та Flask.

Висновки

В результаті було розроблено інформаційну технологію та програмне забезпечення класифікації банківських текстів на основі згорткової нейронної мережі, яке порівняно з аналогом має кращу на 7,2% достовірність класифікації банківських текстів. Таким чином, мета роботи досягнута – достовірність класифікації банківських текстів підвищена. У подальшому планується використовувати для класифікації банківських текстів спайкінгові нейронні мережі [3]. Вони більш пристосовані для обробки динамічних образів, ніж класичні нейронні мережі. Крім того, спайкінгові нейронні мережі мають гарні перспективи для апаратної реалізації [4] та найкраще підходять для побудови операційного ядра нейрокомп'ютерів [5].

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. GloVe: Global Vectors for Word Representation – [Електронний ресурс]. – Режим доступу: https://nlp.stanford.edu/pubs/glove.pdf

2. Згорткові нейронні мережі – [Електронний ресурс]. – Режим доступу: http://ru.datasides.com/code/cnn-convolutional-neural-networks/

3. В.Ф.Бардаченко, О.К.Колесницький, С.А.Василецький. Перспективи застосування імпульсних нейронних мереж з таймерним представленням інформації для розпізнавання динамічних образів// УСіМ.-2003-№6.- С. 73-82.

4. Колесницький О. К. Аналітичний огляд апаратних реалізацій спайкових нейронних мереж / О. К. Колесницький // Математичні машини і системи. – 2015. – №1, С.3-19. ISSN 1028-9763 [Електронний ресурс]. Режим доступу - http://www.immsp.kiev.ua/publications/articles/2015/2015_1/01_2015_Kolesnytskyy.pdf

5. Колесницький О. К. Принципи побудови архітектури спайкових нейрокомп'ютерів / О. К. Колесницький // Вісник Вінницького політехнічного інституту. – Вінниця: УНІВЕРСУМ-Вінниця. – 2014. – №4 (115), С.70-78. [Електронний ресурс]. Режим доступу - http://visnyk.vntu.edu.ua/article/view/3697/5416

Перероdов Артемій Олексійович— студент групи 2КН-18м, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: artemtool@gmail.com

Наукові керівники: ***Колесницький Олег Костянтинович*** — к.т.н., доцент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.

Денисов Ігор Костянтинович — викладач кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця

Pererodov Artemiy O. — Department of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, email : artemtool@gmail.com

Supervisors: ***Oleh K. Kolesnytskyj*** — Cand. Sc. (Eng), Assistant Professor, Assistant Professor of the Chair of Computer Science, Vinnytsia National Technical University, Vinnytsia.

Ihor K. Denisov — lecturer of the Chair of Computer Science, Vinnytsia National Technical University, Vinnytsia.