УДК 004.9

РОЗРОБКА ПІДСИСТЕМИ ПІДБОРУ КАТЕГОРІЙ ВАКАНСІЙ ЗА ТЕКСТОВИМ ОПИСОМ НАВИЧОК ІТ-ФАХІВЦЯ

Нікітін Павло Якович,
асистент Петросюк Денис Валерійович
Національний університет «Одеська політехніка», УКРАЇНА

АНОТАЦІЯ. Досліджено існуючі підходи до багатокласової класифікації (з кількома мітками) текстів та визначено переваги методу з використанням TF-IDF. Розроблена підсистема являє собою модель, яка повертає категорії вакансій за введеними навичками користувача. Модель реалізовано за допомогою мови програмування Python та бібліотеки sklearn. Класифікація відбувається з використанням TF-IDF, алгоритму Linear SVC та класифікатору one-vs-rest. Розроблена модель може знайти застосування у сфері пошуку праці та сервісах, які надають такий функціонал.

Вступ. В останні десятиліття інформаційна революція призвела до вибухового зростання обсягу доступної нам інформації, особливо у сфері інтернету та соціальних медіа. Однак, отримання корисної інформації з таких величезних обсягів текстових даних є непростим завданням. Багатокласова класифікація тексту є одним із важливих напрямів у галузі комп'ютерних наук, що дозволяє автоматизовано організовувати, категоризувати та обробляти текстову інформацію.

Багатокласова класифікація тексту має широкий спектр застосувань у різних галузях, включаючи природну мову, інформаційний пошук, соціальні медіа, медицину, фінанси та багато інших. Вона дозволяє визначати категорію або мітку, до якої належить цей текст, ґрунтуючись на його змісті та семантичній структурі. Це відкриває можливості для автоматичної обробки тексту, виявлення тематичних трендів, моніторингу громадської думки, фільтрації спаму та багато іншого.

Працевлаштування стає дедалі актуальнішою проблемою. З кожним роком кількість безробітних зростає, а конкуренція на ринку праці лише посилюється. Щоб знайти роботу, необхідно витратити багато часу та сил на пошук вакансій, перегляд оголошень та складання резюме. Однак з появою мобільних пристроїв та додатків для них цей процес може бути значно спрощений.

Розробка підсистеми, яка допоможе користувачам швидко та зручно знаходити вакансії за їх навичками значно спростить проблему пошуку праці. Також, ця підсистема може бути корисною для бізнесу, який надає сервіси для пошуку праці та бажає надати користувачу функціонал, який якісно відрізнятиме цей сервіс від інших.

Мета роботи. Дослідження існуючих підходів до багатокласової класифікації (з кількома мітками) текстів та на їх основі розробка моделі, яка буде забезпечувати оптимальну точність та високу швидкість класифікації, що дозволить користувачам отримувати відповідні вакансії за їх навичками.

Основна частина роботи. Отримання категорій вакансій за вхідними навичками – це задача обробки природної мови(NLP), а саме класифікація з кількома мітками (multi-label classification). Класифікація з кількома мітками - це задача машинного навчання, що полягає у прогнозуванні належності об'єкта до однієї або більше міток(класів), які визначені заздалегідь.

Для вирішення задачі багатокласової класифікації можуть використовуватись різні алгоритми, такі як логістична регресія, метод опорних векторів, випадкові ліси та нейронні мережі. Навчання моделі включає етапи передобробки даних, такі як масштабування, видалення викидів і обробка категоріальних ознак.

Процес багатокласової класифікації може включати вирішення проблеми незбалансованих класів, коли деякі класи представлені набагато меншою кількістю об'єктів.

Оцінка якості моделі може проводитись з використанням різних метрик, таких як точність, повнота, F-міра та матриця помилок. Остаточне використання моделі багатокласової класифікації включає застосування навченої моделі для прогнозування класу нових, невідомих раніше об'єктів.

Для того, щоб модель правильно працювала та надавала адекватні результати, необхідно попередньо обробити дані. Текстові дані, в даному випадку описи вакансій, мають багато змісту, який не має інформаційної користі для задач класифікації, та навить погіршують точність моделі. Це всілякі символи, союзи, прийменники, тощо. Ці слова називають «стоп-слова» та їх необхідно видалити.

Треба видалити усі спеціальні символи (/\|+-,.) з опису. Це робиться для того, щоб слова java та java. інтерпретувалися як одне й те саме слово. Однак, в описах вакансій будуть зустрічатися такі послідовності, які мають дуже великий інформаційний зміст – c++, c#, .net, тому їх потрібно залишити. Також треба привести текст до нижнього регістру. Це робиться для того, щоб слова Java та java інтерпретувалися як одне й те саме слово

Для розв’язання задачі була обрана модель з використанням TF-IDF, бо у вакансіях є багато слів, які будуть з’являтися у багатьох категоріях, наприклад: кандидат, ООП, алгоритми, тощо. А також є слова, які зустрічаються тільки у конкретних категоріях: laravel, django, hibernate, тощо. Тому, ліпше використовувати модель з TF-IDF, бо в цілому розповсюджені слова будуть мати менший вплив на результат, а конкретні слова, які розповсюджені у конкретних категоріях – більший.

Існує декілька підходів, за якими може відбуватися класифікація текстів. "Мішок слів" (Bag of words, BOW) – це уявлення, що використовується для обробки природної мови У цій моделі текст подається у вигляді мішка (множини) слів, без урахування граматики і порядку слів, але із збереженням кратності. У цьому підході текст сприймається як невпорядкований набір слів. Спочатку створюється словник, що містить усі унікальні слова в навчальному наборі текстів. Потім кожному тексту надається вектор, де кожна компонента відповідає наявності або відсутності певного слова зі словника. Цей підхід ігнорує порядок слів і вважає, що важливою є лише їх наявність.

TF-IDF (term frequency-inverse document frequency) – це числова статистика, покликана відобразити, наскільки важливим є слово для документа в колекції або корпусі. Вона часто використовується як ваговий коефіцієнт при пошуку в інформаційному пошуку та текстовому аналізі. Значення TF-IDF збільшується пропорційно до кількості разів, коли слово з'являється в документі, і компенсується кількістю документів у корпусі, що містять це слово, що допомагає скоригувати той факт, що деякі слова взагалі з'являються частіше. Ця модель є розширенням методу "мішок слів".

У набору даних, на якому буде навчатися та тестуватися модель присутня проблема незбалансованості даних. Незбалансованість – кількісна переваги одного класу над іншими, або наявність класу, зі значно меншою кількістю екземплярів. У нашому випадку присутні обидва варіанта. Проблема незбалансованих даних може мати негативний вплив на процес навчання та якість моделі класифікації. Для вирішення цієї проблеми був обран підхід Oversampling, а саме генерування синтетичних прикладів на основі вже існуючих (SMOTE).

Для рішення задачі багатокласової класифікації були обрані 4 алгоритми. Їх назви та міра F-1, за якою відбувалася оцінка якості моделі (табл. 1).

Таблиця 1 – Таблиця порівняння якості класифікації різних алгоритмів

Алгоритм	F-1 score
Логістична регресія	0.85
Linear Support Vector Machine	0.96
Випадковий ліс дерев	0.92
Стохастичний градієнтний спуск	0.91

За метрикою F1 найліпший класифікатор – Linear SVC. Можна налаштувати його гіперпараметри, щоб отримати покращення результатів. Була використана регуляризація L2 та C = 2. Після налаштування, метрика F-1 підвищилася до 0.97.

Висновки. В результаті була розроблена модель, яка повертає категорії вакансій за введеними навичками користувача з використанням TF-IDF. Такий метод багатокласової класифікації текстів показав ефективність в аналізі великої кількості вакансій та є простим у реалізації.

Реалізована методика багатокласової класифікації текстів вакансій забезпечує точність на рівні 97% та навчена модель швидко повертає дані на рівні <1с, що дозволить користувачам швидко отримувати вакансії за категоріями. Точність класифікації вдалося значно підвищити більш ніж на 40% за рахунок балансування даних та налаштування гіперпараметрів класифікатора.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Multi-Label Text Classification on Stack Overflow Tag Prediction [Електронний ресурс] - https://kgptalkie.com/multi-label-text-classification-on-stack-overflow-tag-prediction/
2. NLP Tutorial - Text Representation: TF-IDF [Електронний ресурс] - https://github.com/codebasics/nlp-tutorials/blob/main/12_tf_idf/tf_idf_tutorial_nlp_codebasics.ipynb/
3. Scikit-learn User Guide [Електронний ресурс] - https://scikit-learn.org/stable/user_guide.html

DEVELOPMENT OF A SUBSYSTEM FOR THE SELECTION OF JOB CATEGORIES ACCORDING TO THE TEXT DESCRIPTION OF THE SKILLS OF AN IT SPECIALIST

Pavlo Nikitin,
assistant Denys Petrosiuk
Odessa National Polytechnic University, Ukraine

ANNOTATION. The existing approaches to multi-class classification (with multi-labels) of texts were studied and the advantages of the method using TF-IDF were determined. The developed subsystem is a model that returns job categories based on the user's entered skills. The model is implemented using the Python programming language and the sklearn library. Classification takes place using TF-IDF, Linear SVC algorithm and one-vs-rest classifier. The developed model can find application in the field of job search and services that provide such functionality.