

Магістрант Коваленко М. В., к.т.н, доцент Юрчишин В. Я.

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

МЕТОД АНАЛІЗУ ТЕКСТОВИХ ОГОЛОШЕНЬ ДЛЯ ПРОГНОЗУВАННЯ ПОПУЛЯРНИХ ВАКАНСІЙ

Abstract

Kovalenko Maksym, student; Yurchyshyn Vasyl, assoc. prof., PhD

Method for analyzing textual advertisements to predict popular job vacancies

This paper presents a hybrid approach to predicting job vacancy popularity by combining natural language processing and machine learning methods. The study explores transformer-based models, particularly BERT, for semantic text analysis integrated with behavioral and metadata features.

Вступ

Сучасний ринок праці характеризується стрімким зростанням онлайн-платформ для пошуку роботи, де щоденно публікуються тисячі текстових оголошень про вакансії. Ефективне прогнозування популярності таких оголошень є важливою задачею як для роботодавців, які прагнуть залучити найкращих кандидатів, так і для платформ, що намагаються оптимізувати рекомендаційні системи та підвищити залученість користувачів.

Традиційні статистичні методи прогнозування, такі як ARIMA та експоненціальне згладжування, базуються виключно на часових рядах і не враховують семантичний вміст текстових оголошень. Водночас, поява методів обробки природної мови та глибокого навчання, зокрема трансформерних моделей на кшталт BERT, відкрила нові можливості для аналізу текстових даних та виявлення складних залежностей між змістом вакансії та її привабливістю для шукачів роботи.

Дослідження в галузі NLP-based рекрутингу демонструють ефективність використання векторних представлень тексту для класифікації та ранжування вакансій. Проте більшість існуючих рішень фокусуються на рекомендаційних системах типу "кандидат-вакансія", тоді як задача прогнозування популярності оголошення ще до його публікації залишається недостатньо дослідженою, особливо в контексті українського ринку праці з його мовною та культурною специфікою.

Невирішеними залишаються питання ефективного поєднання текстових ознак з поведінковими та метаданими, адаптації багатомовних моделей до українських текстів, а також побудови систем реального часу, здатних швидко адаптуватися до динамічних змін ринкових тенденцій.

Постановка задачі

Метою дослідження є розробка методу для прогнозування популярності вакансій на основі комплексного аналізу текстових оголошень із використанням трансформерних моделей обробки природної мови та додаткових числових ознак.

Задача формалізується як бінарна класифікація: за текстом вакансії та супутніми характеристиками визначити, чи буде вакансія популярною (кількість переглядів або відгуків перевищує заданий поріг за певний період). Альтернативно розглядається постановка задачі регресії для передбачення конкретного числового значення метрики популярності.

Термінологія

NLP (Natural Language Processing, обробка природної мови) – галузь штучного інтелекту, що займається взаємодією між комп'ютерами та людською мовою, зокрема аналізом, розумінням та генерацією природномовних текстів [1].

BERT (Bidirectional Encoder Representations from Transformers) – трансформерна модель глибокого навчання для обробки природної мови, яка використовує двонаправлений контекст для розуміння значення слів у реченні [2].

Attention mechanism (механізм уваги) – компонент нейронної мережі, що дозволяє моделі зосереджуватися на найбільш релевантних частинах вхідних даних при обробці інформації [3].

Fine-tuning (донавчання) – процес адаптації попередньо навченої моделі до специфічної задачі шляхом додаткового навчання на доменних даних.

TF-IDF (Term Frequency-Inverse Document Frequency) – статистична міра, що визначає важливість слова в документі відносно колекції документів, враховуючи частоту його використання [4].

Токенізація – процес розбиття тексту на окремі елементи (токени), якими можуть бути слова, підслова або символи, для подальшої обробки.

Вибір методів та інструментів

Для розв'язання поставленої задачі обрано комбінований підхід, що поєднує сучасні методи обробки природної мови з класичними алгоритмами машинного навчання. Основні етапи обробки включають попередню

обробку тексту, векторизацію, побудову моделі класифікації та інтеграцію додаткових ознак.

Попередня обробка текстових даних здійснюється за допомогою бібліотек spaCy та NLTK і включає токенізацію, нормалізацію (приведення до нижнього регістру), видалення HTML-тегів та спеціальних символів, лематизацію та фільтрацію стоп-слів. Для україномовних текстів використовуються спеціалізовані моделі uk_core_news або stanza.

Для векторизації тексту розглядаються три основні підходи. Метод TF-IDF забезпечує базову векторизацію на основі частоти слів, проте не враховує контекст та семантичні зв'язки. Моделі Word2Vec та FastText створюють векторні представлення слів у багатовимірному просторі, причому FastText краще працює з морфологічно складними мовами завдяки врахуванню підслівних одиниць. Найперспективнішим є використання трансформерної моделі BERT, зокрема багатомовної версії multilingual BERT або україномовної UKRBERT, що дозволяє враховувати контекст слів та виявляти семантичні залежності завдяки механізму attention.

Для побудови класифікаційної моделі порівнюються декілька алгоритмів. Логістична регресія слугує базовим інтерпретованим методом для порівняння. Ансамблеві методи Random Forest та градієнтний бустинг (XGBoost, LightGBM) демонструють високу точність на структурованих даних. Багатошаровий перцептрон (MLP) використовується як проміжний варіант між класичними методами та глибокими мережами. Основною моделлю є BERT із додатковим класифікаційним шаром, що проходить процес fine-tuning на доменних даних.

Ключовою особливістю запропонованого методу є інтеграція текстових ознак з додатковими числовими характеристиками: середня популярність аналогічних вакансій за останній період, рівень конкуренції в категорії, наявність ключових слів (віддалена робота, рівень досвіду), день тижня та час публікації, метадані про компанію та локацію. Ці ознаки подаються як окремі входи до моделі або інтегруються у повнозв'язний шар після BERT-кодувальника.

Практичне застосування

Розроблене програмне забезпечення може використовуватися платформами пошуку роботи для автоматичної оцінки якості оголошень та надання рекомендацій роботодавцям щодо покращення тексту вакансії. HR-відділи компаній можуть застосовувати систему для попередньої оцінки очікуваного відгуку перед публікацією вакансії та оптимізації формулювань для підвищення залученості кандидатів.

Аналітичні сервіси можуть інтегрувати модель для дослідження трендів ринку праці та виявлення найпривабливіших характеристик

вакансій в різних галузях. Рекрутингові агентства отримують інструмент для швидкої оцінки конкурентоспроможності пропозицій клієнтів та формування стратегій залучення талантів.

Запропонований спосіб є корисним для задач різного масштабу, де необхідний баланс між швидкістю та точністю прогнозування. Його архітектурна гнучкість дозволяє застосовувати його в різних практичних контекстах, де рішення мають прийматися швидко і без залучення значних обчислювальних ресурсів. Модульна структура програмного забезпечення забезпечує можливість поетапного впровадження та адаптації до специфічних вимог бізнесу.

Висновки

У роботі запропоновано та реалізовано метод прогнозування популярності вакансій, що базується на комплексному аналізі текстових оголошень із використанням трансформерної моделі BERT та додаткових числових ознак.

Ключовими перевагами розробленого рішення є здатність враховувати семантичний контекст тексту завдяки трансформерній архітектурі, можливість інтеграції різнорідних ознак у єдину модель, адаптованість до специфіки українського ринку праці та масштабованість завдяки мікросервісній архітектурі.

Подальші дослідження будуть спрямовані на розробку мультимодального підходу з урахуванням зображень (логотипи компаній, інфографіка), впровадження механізму самонавчання моделі на основі зворотного зв'язку від користувачів, розширення системи для роботи з багатомовними вакансіями та адаптацію моделі для прогнозування інших метрик успішності (швидкість закриття вакансії, якість кандидатів).

Література

1. Natural language processing – [Електронний ресурс] – Режим доступу: https://en.wikipedia.org/wiki/Natural_language_processing
2. BERT (language model) – [Електронний ресурс] – Режим доступу: [https://en.wikipedia.org/wiki/BERT_\(language_model\)](https://en.wikipedia.org/wiki/BERT_(language_model))
3. Attention (machine learning) – [Електронний ресурс] – Режим доступу: [https://en.wikipedia.org/wiki/Attention_\(machine_learning\)](https://en.wikipedia.org/wiki/Attention_(machine_learning))
4. Tf-idf – [Електронний ресурс] – Режим доступу: <https://en.wikipedia.org/wiki/Tf%E2%80%93idf>