МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Сумський державний університет

Факультет електроніки та інформаційних технологій

Кафедра комп'ютерних наук

«До захисту допущено»

В.о. завідувача кафедри

_____ Ігор ШЕЛЕХОВ

(підпис)

09 червня 2024 р.

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня бакалавр

зі спеціальності 122 – Комп'ютерних наук,

освітньо- професійної програми «Інформатика»

на тему: «ІНФОРМАЦІЙНА СИСТЕМА ОЦІНКИ ПРОФЕСІЙНОЇ ВІДПОВІДНОСТІ КАНДИДАТА»

здобувача групи Ин-04р, Сапаргул Рахманова

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_____ Сапаргул Рахманова

(підпис)

Керівник

доцент,

кандидат технічних наук

_____

(підпис)

Суми – 2024

Сумський державний університет
Факультет електроніки та інформаційних технологій
Кафедра комп'ютерних наук

«Затверджую»
В.о. завідувача кафедри
_____ Ігор ШЕЛЕХОВ
(підпис)

ЗАВДАННЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня бакалавра

зі спеціальності 122 - Комп'ютерних наук, освітньо-професійної програми «Інформатика»
здобувача групи Ин-04р, Сапаргул Рахманова

1. Тема роботи: «інформаційна система оцінки професійної відповідності кандидата»
затверджую наказом по СумДУ від *«01» червня 2024 р. № 0475-VI*
2. Термін здачі здобувачем кваліфікаційної роботи *до 09 червня 2024 року*
3. Вхідні дані до кваліфікаційної роботи _____
4. Зміст розрахунково-пояснювальної записки (перелік питань, що їх належить розробити)
1) Огляд сучасного стану застосування інтелектуальних систем при перевірці відповідності кандидата 2)Огляд методів психологічної оцінки кандидата. 3) Формування та аналіз вхідних даних. 4) Аналіз статистичних метрик при роботі з природньою мовою 5) Реалізація алгоритму класифікації для оцінки резюме
5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)_____
6. Консультанти до проєкту (роботи), із значенням розділів проєкту, що стосується їх

Розділ	Консультант	Підпис, дата	
		Завдання видав	Завдання прийняв

7. Дата видачі завдання «____» _____ 20 ___ р.

Завдання прийняв до виконання _____ Керівник _____
(підпис) (підпис)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Назва етапів кваліфікаційної роботи	Термін виконання	Примітка
1	*Огляд методів психологічної оцінки кандидата*		
2	*Аналіз статистичних метрик при роботі з природньою мовою*		
3	*Огляд сучасного стану застосування інтелектуальних систем при перевірці відповідності кандидата*		
4	*Формування та аналіз вхідних даних*		
5	*Реалізація алгоритму класифікації для оцінки резюме*		

Здобувач вищої освіти _____ Керівник _____
(підпис) (підпис)

АНОТАЦІЯ

Записка: 41 стр., 9 рис., 3 таблиці, 1 додаток, 19 використаних джерел.

Обґрунтування актуальності теми роботи – Було розроблено автоматизовану систему підтримки прийняття рішень для оцінки релевантності навичок та вмінь кандидата щодо вимог професії. Таким чином користувач може отримати рекомендації при виборі місця роботи зарахунок автоматичного статистичного аналізу резюме та співставлення його з апріорно отриманими даними.

Об’єкт дослідження – процес підтримки прийняття рішень.

Мета роботи – розробка системи автоматизованої оцінки відповідності кандидата заданим професіям.

Методи дослідження – моделі та методи інтелектуального аналізу резюме кандидата.

Результати – автоматизована система підтримки прийняття рішень для оцінки релевантності навичок та вмінь кандидата щодо вимог професії.

NLP, ІНТЕЛЕКТУАЛЬНА СИСТЕМА, ІНФОРМАЦІЙНА ПІДТРИМКА, МАШИННЕ НАВЧАННЯ, СЕМАНТИЧНИЙ АНАЛІЗ, КЛАСИФІКАЦІЯ

ЗМІСТ

ВСТУП

Кожна людина, свідомо чи ні, у своєму житті робить вибір на користь того чи іншого виду діяльності. Обираючи професію, хобі чи вид творчої самореалізації. У деякому сенсі цей вибір – є інвестицією часу. Одного з не багатьох невідновлюваних ресурсів.

Хоча найочевиднішим методом обрання професії є власні вподобання, проте мало хто з людей, в повній мірі, здатен оцінити свої можливості у новій для себе справі. До того ж спробувати усі існуючі варіанти, здебільшого, не можливо. Тим не менш, кожен вид діяльності піддається узагальненню та формалізації. У такий спосіб можливо розгледіти головні вимоги до виконавця. Не очевидні психологічні риси та схильності якими він повинен володіти.

Такий підхід до вівісекції видів професій дозволяє допомогти людині обрати найбільш відповідну для неї роботу. Врахувати її психологічні схильності та зіставити їх з професією у якій би вони були найбільш ефективними.

1 ІНФОРМАЦІЙНИЙ ОГЛЯД

1.1 Сучасний стан застосування інтелектуальних систем при перевірці відповідності кандидата.

З розвитком алгоритмів штучного інтелекту (ШІ) з’явилася чи мала кількість програмних застосунків, які автоматизують процес роботи з резюме. Зазвичай вони зосереджені або на оцінці відповідності резюме кандидата [1] робочій позиції на яку він претендує, або ж на безпосередньому створені відповідного документу [2, 3].

У першому випадку додатки використовуються менеджерами по персоналу для відсіювання кандидатів. Тим не менш, вони враховують лише змістовну частину резюме, що дозволяє оцінити тільки технічну відповідність людини конкретному робочому місцю. На практиці, навіть якщо кандидат і володіє усім переліком необхідних навичок, це не означає, що він здатний виконувати поставлену перед ним роботу. Його психологічні схильності здатні ускладнювати, а у деяких випадках й унеможливити робочий процес. Роботодавцю, у більшості випадків, не потрібна людина, яка найкраще знає предметну область своєї професії, йому потрібен той, хто здатний показувати результати в конкретній робочій екосистемі.

У свою чергу, коли кандидат використовує програмні застосунки для написання власного резюме, то він зосереджений на його змістовній частині. Спробам створити такий текст, який би найкраще, з технічної точки зору, відповідав конкретній пропозиції на ринку праці. При цьому, мало коли підіймається питання наскільки людині підходить та чи інша професія. І чи здатна вона в повній мірі розкрити її потенціал.

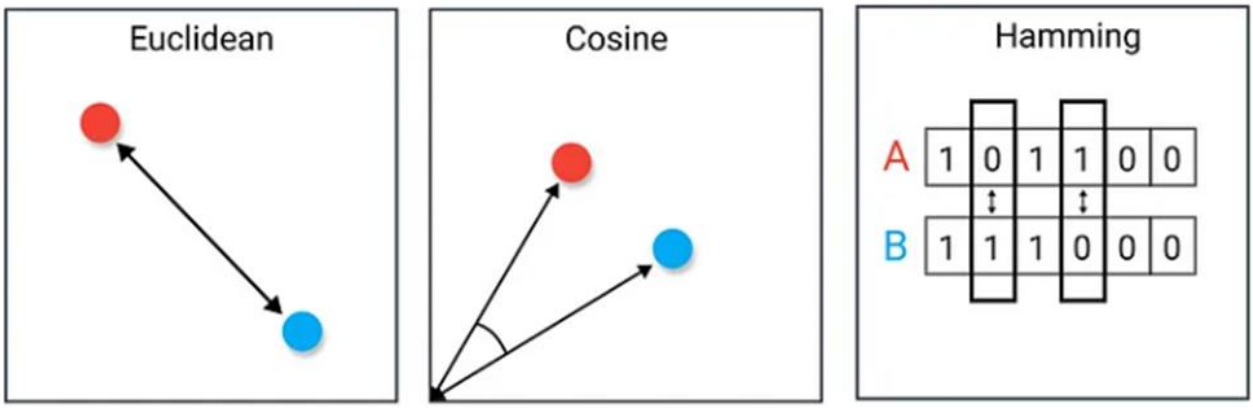
Найбільш слабке місце будь-якого бізнесу – це людина. Необхідно приймати на роботу таких співробітників, які здатні надати максимальну користь протягом тривалого часу. Для цього важливим критерієм є психологічний портрет людини. Хоча у відкритих джерелах можливо знайти чи малу кількість інформації про різні професії, наприклад [4]. Проте, у більшості випадках, це

лише їх технічний опис. Умови, темп та вимоги до конкретного місця роботи завжди будуть різні. Тому навіть, якщо людина вміє все з переліку вимог до її навичок, вона далеко не завжди може працювати у конкретній робочій екосистемі.

Звісно, що існують програмні забезпечення, які здатні давати рекомендації стосовно працевлаштування [5]. Такі алгоритми використовують інформацію про навички, досвід та освіту кандидата порівнюючи їх з величезною базою даних відомих вакансій та їх вимог. У деяких випадках ці інструменти навіть дозволяють спрогнозувати майбутній кар'єрний шлях. Проте, цього не достатньо, коли потрібно дізнатися чи підходить конкретна фірма кандидату, чи здатний він бути в ній максимально корисним.

На етапі проектування інтелектуальної компоненти подібного типу варто зауважити, що насамперед вхідними даними для неї можуть бути як текстові дані, так й числові. У першому випадку обробці буде підлягатися безпосередній текст резюме. У другому може бути використаний попередньо оброблений матеріал. Наприклад, результат опитування або ж тестування. У залежності від вхідних даних буде вибраний простір ознак у якому відбуватиметься кінцева класифікація об'єкту інтересу.

Дистанційні метрики, які формують робочий простір ознак, є не однаково ефективними для різних вхідних даних. Існує приблизно 9 основних способів (рис.1.1) обчислення відстані між векторами ознак. Головна логіка полягає в тому, що чим ближче знаходяться між собою точки в абстрактному просторі, тим вони більш схожі. Тим не менш, фундаментальною проблемою тут є визначення міри схожості, а отже близькості. Адже простір даних абстрактний, а отже і дані можуть в ньому розміщуватися довільним чином. Це залежить не тільки від їх природи, а й від процесу обробки. Тому можна стверджувати, що цифрові дані, як й реальні, є суб'єктивними і їх правильна інтерпретація залежить від багатьох зовнішніх факторів.



а б в

Рисунок 1.1. Приклади дистанційних метрик: а) евклідова відстань; б) кутова відстань; в) бінарна відстань Хеммінга

Найпростіша метрика, з якою знайомий кожен школяр, це евклідова відстань, яка формує однойменний простір ознак. По своїй суті, це найкоротший відрізок, який поєднує дві довільні точки (рис.1.1.а) і визначається згідно (1.1).

$$d(p,q) = d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{1.1}$$

При аналізі існуючих фреймворків машинного навчання можна помітити, що більшість алгоритмів інтелектуального аналізу використовують саме цю метрику, наприклад, в кластер аналізі за методом K-найближчих сусідів. Проте евклідова відстань не найкращим чином себе проявляє при малій кількості даних та при використанні векторів ознак надмірної розмірності. Чим більше даних і чим менше вектор, тим кращий результат класифікації об’єктів в евклідовому просторі.

Косинусова метрика (рис 1.1.б) – є логічним доповненням до евклідового простору. У рамках якого відстань визначається як кут між двома векторами (1.2).

$$A \cdot B = \|A\| \, \|B\| \cos(\theta) \tag{1.2}$$

При такій реалізації коефіцієнт подібності визначений [-1, 1], де -1 – це абсолютно протилежні між собою ознаки, а отже і не схожі; 0 – відсутня кореляція, але можлива подібність; 1 – це повна схожість двох ознак. Візуально це достатньо легко пояснити (рис.1.2), адже якщо два вектори протилежні між собою, мають кут 180º, то косинус між ними буде дорівнювати -1.

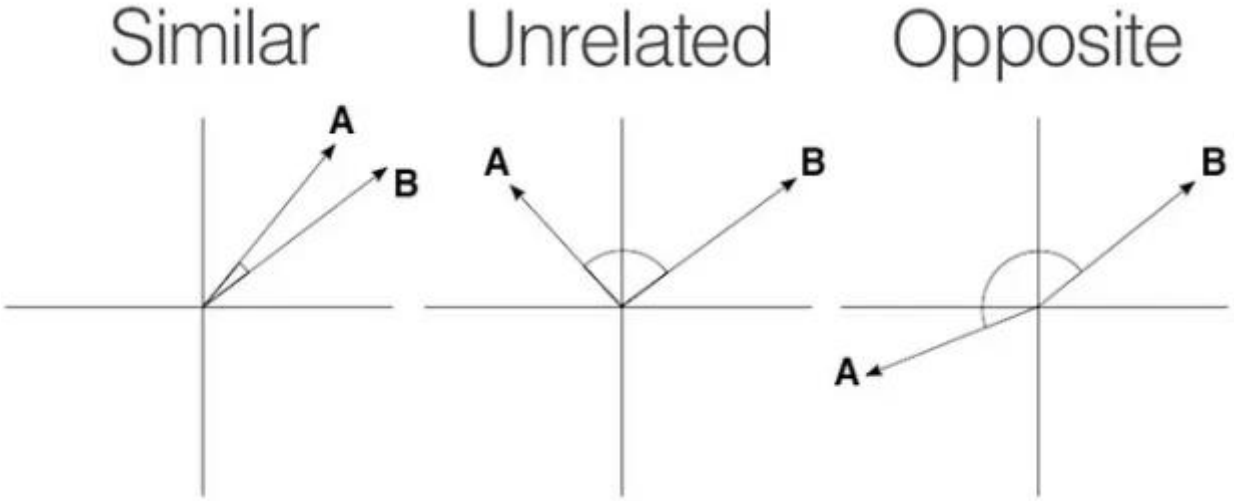


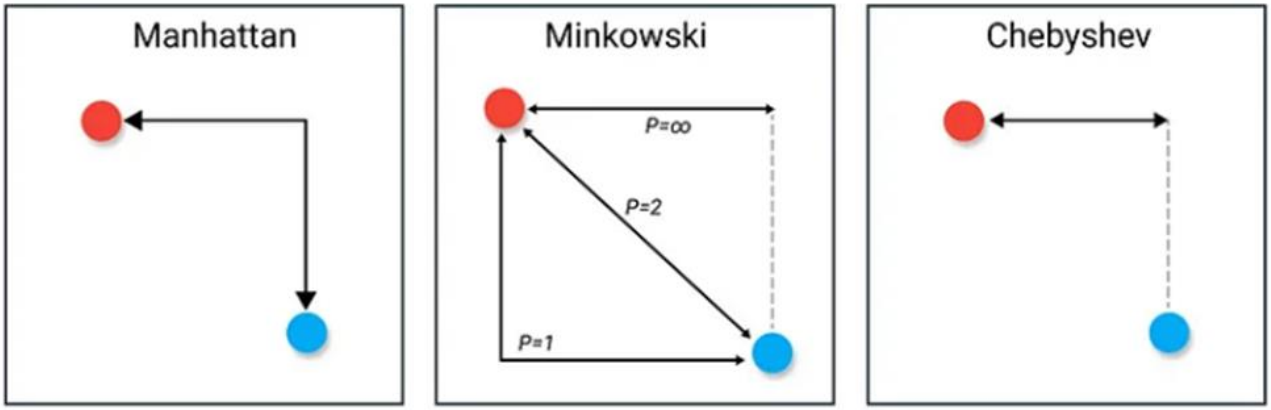
Рисунок 1.2. Приклад схожості векторів-ознак за косинусовою метрикою

На практиці ця метрика використовується наступним чином: кут між векторами показує наскільки схожі між собою вектори, а відстань показує наскільки вони відрізняються. Іноді можна зустріти поєднання цієї метрики з евклідовою відстанню, але у більшості випадках використання косинусу як основного коефіцієнту подібності двох векторів може бути обумовлено лише вхідними даними та способом їх нормалізації.

З точки зору програмування, найбільш очевидною дистанційною метрикою для цифрових даних є відстань Хеммінга (рис.1.1.в), яка полягає у порівнянні двох бінарних векторів. Таким чином, відстань між об’єктами в бінарному просторі Хеммінга визначається різницею бітів їх векторів.

Теоретично метрика Хемінга є найбільш універсальним способом побудуванням n-мірного простору ознак, проте на практиці виникає не очевидна проблема бінаризації даних. Не можливо однозначно сказати, як краще це

зробити для конкретного об’єкта з мінімізацією втрати його початкової інформативності. Цей вибір є максимально суб’єктивним і залежить від інженера. Тому у більшості фреймворків метрика Хеммінга не використовується на пряму при машинному навчанні.



а б в

Рисунок 1.3. Приклади дистанційних метрик: а) Манхтеенська відстань; б) Мінховська метрика; в) відстань Чебішеева

Спрощеною альтернативою еевклідової метрики є Манхетенська відстань (рис.1.3.а), яка визначається наступним чином (1.3):

$$D = \sum_{i=1}^{n} |x_i - y_i| \tag{1.3}$$

Головною її перевагою є відсутність зведення в квадрат та обчислення коренів, які присутні в (1.1), що робить простір ознак оснований на Манхетенській метриці більш зручним для масштабування, тому цій метриці віддають перевагу при роботі з багатовимірними векторами, при цьому, на відмінну від простору Хеммінга, не посилюється різниця й не ігноруються особливості між окремими ознаками.

У свою чергу Мінховська метрика (рис.1.3.б) є поєднанням Евклідової та Манхетенської, а її головною перевагою є те, що вона враховує просторове

положення довільних векторів. У свою чергу це додає більш повне уявлення про їх схожість у просторі ознак. Мінховська відстань визначається наступним чином (1.4):

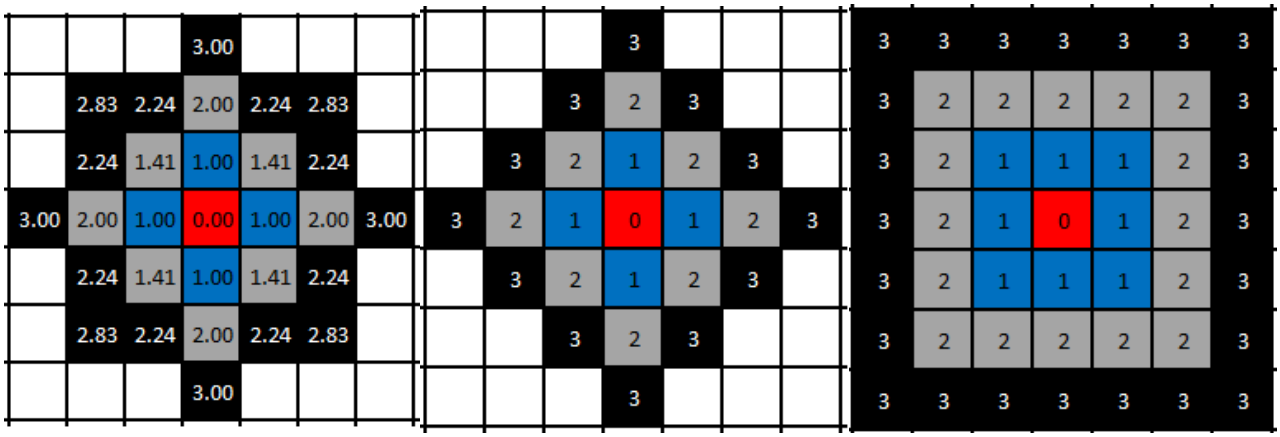
$$\mathrm{d}(p,q) = \left(\sum_{i=1}^{n} |p_i - q_i|^c\right)^{\frac{1}{c}}, \tag{1.4}$$

де c – це порядок норми, при c=1 – формула набуває вигляду (1.3), а при c=2 формула набуває вигляду (1.1)

Достатньо специфічною, в машинному навчанні, є метрика Чебішева (рис 1.3.в), при які знаходиться максимальна відстань між двома довільними точками (1.5):

$$D = \sum_{i=1}^{n} \max(|x_i - y_i|) \tag{1.5}$$

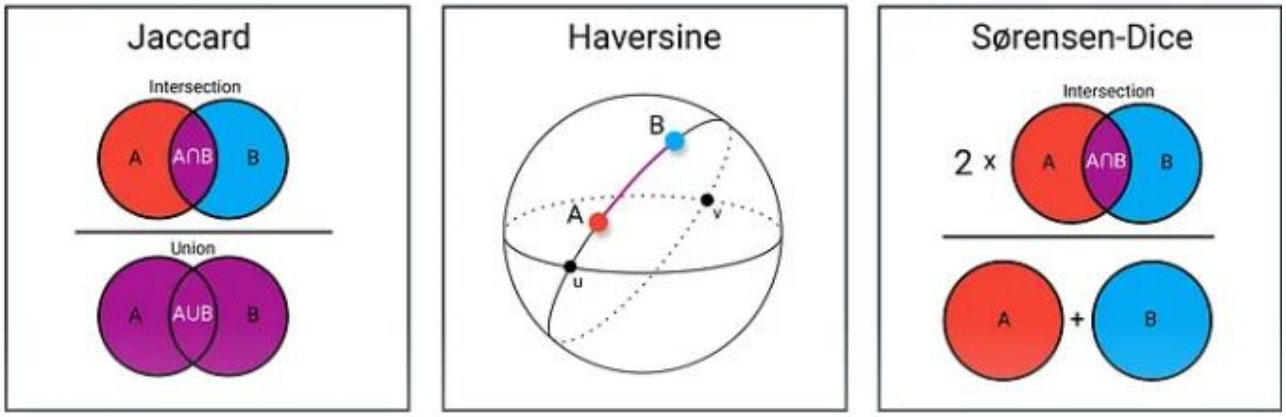
На рисунку 1.4 візуально показано різницю між евклідовою, манхетенською та метрикою чебішева. У вигляді двовимірного простору ознак та відстані центрального елементу до довільної ознаки.



а б в

Рисунок 1.4. Приклади відстаней для різних метри: а) евклідова; б) манхетенська; в) Чебішеева

Припустимо, що чорний, сірий та синій колір – це різні класи розпізнавання, тоді на рисунку 1.4. добре видно різницю при класифікації ознак для різних метрик. У деякому сенсі можна стверджувати, що від обраного способу обчислення відстані залежить ступінь абстракції вирішальних правил, адже, наприклад при застосуванні метрики Чебішева (рис.1.4.в) було отримано кількісно більше представників кожного з класу. Це свідчить про те, що ступінь абстракції об’єкту значно вища. Тому для найдоцільнішого вибору метрик варто проводити детальний аналіз вхідних даних, та експериментально підтверджувати свій вибір.



а б в

Рисунок 1.5. Приклади дистанційних метрик: а) Джакард; б) Гаверсунус; в) індекс Соренса

Метрика Джакарда (рис.1.5.а) є найбільш тривіальною з усіх вже розглянутих і у своїй основі відноситься більше до статистичних методів, ніж до дистанційних. Логіка полягає у тому, що для знаходження подібності двох об’єктів, у нашому випадку векторів-ознак, відбувається співставлення усіх їх елементів. Чим більше схожих ознак між собою, ти більше вони подібні. Нажаль у такому вигляді не можливо створити повноцінний простір ознак, але мтрика Джакарда дуже часто використовується у модифікаціях алгоритмів машинного навчання. Зокрема із-за своєї простоти та статистичної основи.

Продовженням ідеї статистичного порівняння математичних об’єктів є використання індексу Соренса (рис 1.5.б), який описується наступним чином:

$$QS = \frac{2C}{A+B}, \text{де} \tag{1.6}$$

A і B – це кількість ознак в кожному з відповідних векторів; C – число ознак, які є спільними для двох векторів.

Ця статистична метрика є гарним прикладом специфічного коефіцієнту схожості, адже індекс Соренса найчастіше використовують в природознавстві та екології. Знання про такі статистичні покажчики є вкрай важливі при розробці вузькоспеціалізованого програмного забезпечення, де на вхід подаються одноманітні та добре досліджені дані.

Окрім цього вузькоспеціалізованим способом створення простору ознак – є формула Гаврсунуса (рис. 1.5.b), яка була спеціально розроблена для картографування і активно використовується в Google Maps API [7]. Ця метрична система дозволяє обчислювати відстань між точка на місцевості, які задаються через широту та довготу, враховуючи масштаб. Для цього враховується кривизна Землі при вводі вхідних даних. Сама ж формула Гаверсунуса виражається череез три величини (1.7):

$$\begin{cases} a = sin^2\left(\phi B - \frac{\phi A}{2}\right) + \cos(\phi A) * \cos(\phi B) * sin^2\left(\lambda B - \frac{\lambda A}{2}\right) \\ c = 2 * atan2(\sqrt{a}, \sqrt{1-a}) \\ d = R * c \end{cases}, \text{де} \tag{1.7}$$

A і B – це деякі точки, що задаються через широту та довготу; ϕ – широта в радіанах; λ – довгота в радіанах; R – це приблизний радіус планети, щр дорівнює приблизно 6371.04 кілометрів.

Очевидно, що не вдасться примітити формулу Гаверсунус в кожній прикладній задачі, але це доцільний приклад того, що існують сфери людської діяльності, де не має потреби в створенні абстрактного простору ознак, адже він вже існує для цих даних. При цьому це стосується не тільки просторових даних.

Звісно, що було розглянуто далеко не всі існуючі метрики, але усі згадані є найбільш популярними та часто застосовуваними в найрізноманітніших фреймворках для машинного навчання. Можливо, що при атипових вхідних даних, або специфічних варіантах використання кінцевої системи варто звернути увагу й на інші.

Двосторонньо ефективну систему запропонували автори [6], яка пропонує кандидату створити довільне резюме, де він опише усі свої сильні сторони та наддасть перелік технічних навичок. У свою чергу алгоритм знайде найбільш відповідну для нього пропозиції по працевлаштуванню. У випадках, коли підприємство формує вимоги до кандидата, то алгоритм може сформувати характеристику обраної посади, згідно бази даних. Автори стверджують, що імплементація їх програмного застосунку здатна пришвидшити роботу відділу кадрів на 70%.

Структура їх розробки складалася з наступних етапів:

1) На вхід подається резюме кандидата, яке написане природньою мовою та має довільну структуру;
2) Текст аналізується методами обробки природньої мови (NLP), щоб віднести кожне речення до однієї з абстрактних груп:

   Компетенція кандидата – освіта, навички, риси характери про які в явно вигляді було вказано в резюме.

   Поведінкова характеристика – припущення або виводи про наявні психологічні риси кандидата.
3) На виході кандидат отримує перелік найбільш відповідних пропозицій на ринку праці;

Автори зосередили свою увагу на лексичному аналізі резюме. Таким чином, їх система виокремлює слова, що були використані кандидатом. Припускається, що в резюме кандидатів, які намагаються працевлаштуватися за однаковою професією, будуть статистично схожі. Із цього припущення можна

зробити ще один вивід, що такі кандидати будуть схожі між собою за поведінкової характеристикою.

Важливим етапом роботи [6] є опитування кандидатів, яке проводиться після використання розробленої системи. Зокрема чи дійсно обрана системою посада підійшла людині. Оскільки, єдиний адекватний спосіб оцінити подібну систему є прямі відгуки користувачів.

Цікавий поєднання алгоритмів машинного навчання з психологією запропонували в роботі [8]. Автори спробували створити максимально комплексну емоційну мапу ймовірностей реакцій людини при різних тригерах. Насамперед, це корисно, щоб оцінити, а іноді й передбачити реакції співробітника у стресових ситуаціях. Автори припускають, що у більшості випадках люди однаково реагують на типові соціальні взаємодії. Тим не менш, це дослідження може бути цікавим ще з точки зору оцінки кореляції емоцій представників різних професій, або ж рівня заробітку, про що також було згадано в роботі [8].

Отже, сучасні системи інтелектуального аналізу резюме, перш за все, зосереджені на оцінці технічної відповідності наповнення документу, а не на придатності людини до обраного місця роботи. Необхідно розробити систему, яка здатна знайти таку робочу екосистему, де кандидат міг би бути максимально продуктивним.

1.2 Огляд методів психологічної оцінки кандидата

Вивчення професійних інтересів займало психологів упродовж століть. Причини очевидні: по-перше, робота є одним із найважливіших видів діяльності у нашому житті (тобто більшість дорослих витрачають інвестують власний час в робочий процес більше, ніж на сон і розваги); по-друге, навіть в епоху «гнучких працівників» та кар'єрної мобільності більшість людей дотримуються тієї професійної сфери праці, яку вони обирають; по-третє, як і у випадку з романтичними партнерами, більшість людей мають труднощі з вибором роботи

або кар'єри, які їм дійсно подобаються. Таким чином, професійні інтереси є широко прикладною галуззю психології, оскільки вони мають значення для викладачів, роботодавців та консультантів, а також для кожної людини, яка сподівається покращити своє розуміння своєї «професійної придатності».

У «психології вибору кар'єри» є одна цікава особливість, яка полягає в тому, що після 1980-х років дослідження було майже завершене: внесок Джона Холланда, який практично «вбив» дослідження в цій галузі. Незвичайна причина в тому, що його теорія була надто гарною. Він розвив ідею про «таксономію» (класифікаційних сфер) робочого середовища, що ефективно дозволило йому організувати всі існуючі професії (у 60-х, 70-х і 80-х роках) у ширші сімейства посад - так само, як психолог це робить з психологічними якостями людини.

За загальним визнанням, типи посад тоді були набагато простішими і їх було менше, ніж зараз, але таксономія Голландії, як і раніше, дуже актуальна. Якщо детально проаналізувати ринок праці, можна знайти сотні тисяч професій, але всі вони, як і раніше, групуються за основними голландськими «типами» - це реалістичний, дослідницький, художній, соціальний, підприємливий та традиційний. Цікавий аспект цієї таксономії полягає в тому, що її можна застосовувати не лише до профільних професій, а й до робочого середовища, організацій, команд та окремих людей — по суті, це робить теорію Холланда теорією особистості для всього.

Останніми роками тема професійних інтересів знову привернула увагу провідних дослідників особистості. Одна з причин полягає в тому, що хоча модель Холланда добре зарекомендувала себе для профілювання посад, вона насправді не пережила «особистісну війну» між багатьма таксономіями, запропонованими для профілювання людей. Натомість більшість людей думають про особистість у термінах «Великої п'ятірки» (ОКЕАН), тому останніми роками було проведено багато досліджень, присвячених тому, як особистість пов'язана з вибором кар'єри Холланда.

Проблема таксономії полегшується тим фактом, що люди не дуже часто змінюють кар'єру. Дані досліджень близнюків показують, що професійні інтереси – всупереч очікуванням – навіть стабільніші, ніж особистісні характеристики! Дійсно, у людини більше шансів змінити свою особистість, ніж уподобання в роботі, хоча обидва вони далеко не незалежні.

Для оцінки кандидата було використано психологічні методики. Передбачається, що у такий спосіб вдасться максимально комплексно перевірити людину і виявити її приховані особливості та наявні схильності, які можуть стати у нагоді при виконанні професійної діяльності.

- ***Акцент.*** визначає, які риси особистості посилені, які ослаблені, які знаходяться на середньому значенні. Зі поєднання слабких і сильних рис і складається, по суті, наша унікальність і несхожість на інших.
- ***16PF (КЕТЕЛ).*** забезпечує вимірювання особистості і може також використовуватися психологами та іншими фахівцями в галузі психічного здоров'я як клінічний інструмент для діагностики психічних розладів, а також для допомоги в прогнозуванні та терапії. 16PF також може надати інформацію, що відноситься до клінічного процесу та процесу консультування, таку як здатність людини до розуміння, самооцінка, когнітивний стиль, інтерналізація стандартів, відкритість до змін, здатність до емпатії, рівень міжособистісної довіри, якість уподобань, міжособистісні потреби, до влади, реакція на динаміку влади, толерантність до фрустрації та стиль подолання. Таким чином, прилад 16PF дозволяє клініцистам вимірювати тривогу, пристосування, емоційну стабільність та поведінкові проблеми в межах норми. Клініцисти можуть використовувати результати 16PF для визначення ефективних стратегій створення робочого альянсу, розробки терапевтичного плану та вибору ефективних терапевтичних втручань або способів лікування. Його також можна використовувати в інших галузях психології, таких як кар'єра та професійний відбір.

- ***СОНДІ-ТЕСТ.*** Невербальний, проективний, особистісний тест, метою якого є виявлення психічних відхилень. Базується на положенні, що типологічно різні особистісні структури можуть бути представлені поєднаннями 8 основних потягів. Кожен з яких виявляє (за допомогою тесту Сонді) ту чи іншу патологію чи проблему особистості. В обґрунтування свого тесту, Сонді висловлює припущення, що найбільш виражену силу та психодіагностичне значення мають портрети, які відповідають найбільш значним потребам індивіда та відповідають його генетично обумовленим нахилам.

Кожна з наявних методик представляє із себе тестування, результати якого розкривають психологічний портрет людини. Одночасне їх використання дозволить виконати максимально комплексний огляд кандидата.

Варто розуміти, що характер професійної кар'єри в значній мірі суб’єктивний. Кожна людина прагне реалізувати свої власні цілі відповідно до прийнятих цінностей та установок. Важливо усвідомлювати, що уміння сформулювати ціль - є важливою умовою досягнення успіху. Ті, хто розуміють власні сильні сторони відчувають себе більш впевнено, а правильно обрана робоча атмосфера доброякісно впливає на психологічне та фізичне здоров’я людини.

Динаміка соціально-економічних змін спонукає людей до нових викликів та змін у формі діяльності протягом своєї професійної кар’єри. Насамперед, це пов'язано з дефіцитом робочих місць на ринку праці, нових способів і характеру професійних завдань, а також збільшенням вимог щодо професійної кваліфікації. Тому дуже важливо бути гнучким і здатним змінити професійну діяльність відповідно до нових потреб ринку праці.

Одна з небагатьох тем, які психологи ще мають вивчити у зв'язку з професійними інтересами, — це те, чому деякі люди вважають за краще працювати за межами чітко визначеної кар'єри. Адже поточні цифри припускають, що до 40% людей у деякий момент свого життя вирішують

працювати на себе, а деякі назавжди і психологічні причини цього рішення ще достеменно невідомі.

Як довели автори у роботі [9], що найбільшу ефективність робітники демонструють в прийнятній для себе робочій екосистемі. Перш за все, це пов’язано з організацією робочого процесу, який залежить від ідеології компанії. Для забезпечення здорової та продуктивної екосистеми, необхідно поєднати декілька важливих факторів:

- Адміністратори та менеджери повинні однозначно визначитися з методологією управління проектом й ознайомити з внутрішніми правилами кожного нового співробітника. Таким чином, вдасться нормалізувати робочу атмосферу, де кожен учасник розуміє що і в якому вигляді від нього вимагається. Це стосується не стільки вимог до технічних навичок робітника, а внутрішніх, іноді не очевидних, правил, темпу та послідовності при виконанні роботи.
- Кандидат повинен усвідомлювати не формальні вимоги, які виставляє компанію щодо його роботи. Це можуть бути правила поведінки на робочому місці, субординація, темп роботи, графік та інше. Такі вимоги, що не пов’язані з технічними навичками кандидата, але є невід’ємними складовими саме цієї компанії, де кожен співробітники, не зважаючи на свою посаду та роль, виконує їх.
- Специфіка послуг або продукту, який виробляє компанія. Майбутній кандидат повинен розуміти потреби свого кінцевого споживача, адже це впливає на планування його робочого процесу. Таким чином, з’являються не очевидні вимоги до персоналу та задачі для окремих працівників, які можуть бути не пов’язані з технічними навичками кандидата, але бути наслідком специфіки продукту, який виготовляє компанія. Особливо це стосується невеликого бізнесу, де, із-за браку персоналу, делегація обов’язків є достатньо розмитою.

Зазначені фактори формують, так звану, ідеологію компанії. У таких випадках роботодавець повинен підбирати персонал, який би відповідав заданим, не формальним критеріям. Насамперед це робиться для того, щоб сформувати здорову та продуктивну атмосферу в компанії, де кожен співробітник не відчуває дискомфорту від робочого процесу, який включає в себе не тільки прямі обов'язки працівника.

Одним із способів дослідження ринку праці є створення формалізованого портрету ідеального співробітника в конкретної галузі. Таким чином, у роботі [11] автори намагалися сформувати найбільш відповідний психологічний портрет вчителя. Під час свого дослідження вони виявили наступні риси характеру, якими повинен володіти педагогічний спеціаліст: емоційний інтелект, емпатія, адаптивність, емоційна стійкість, саморефлексія, культурна компетенція, навички ефективного спілкування, співпраця та схильність до командної роботи, управління класом, інклюзивність та вміння інтегрувати нові технології у робочій та навчальний процес. Це все перелік не стільки навичок, які визначають компетенцію учителя в своїй сфері інтересу, а список характеристик, які, на думку авторів, повинні бути притаманні ідеальному учителю.

Важливо помітити, що подібна формалізація професії можлива лише у межах певного проміжку часу, адже вона формується на поточних тенденціях та парадигмах. Наприклад, сучасна педагогіка особистісно-орієнтована та передбачає індивідуальний підхід до кожного окремого учня. Тому ідеальний вчитель майбутнього, перш за все, повинен вміти вирішувати індивідуальні проблеми конкретного вихованця. У свою чергу, це означає не тільки інтелектуальний розвиток, а й культурно-духовний.

Подібні дослідження допомагають визначити найбільш релевантні риси характеру для представників різного роду діяльності. При виборі професії це значно спрощує вибір, адже людина може зіставити власні психологічні риси з

формалізованими портретами й визначити, де її характер проявить себе найбільш продуктивним чином.

Тим не менш, сучасна психологія [12] зосереджена на індивідуальному аналізі пацієнта. Досліджень, які розглядають кореляцію між психологічними рисами та професіями не достатньо для формування чіткої закономірності та створення повноцінних систем підтримки прийняття рішень (СППР) для розв'язання прикладних проблем формування персоналу. Однак це перспективний, прикладний напрям роботу, який дозволяє поєднати інформаційні технології, теорію ефективного керування і психологію для максимізації ефективності бізнесу.

1.3 Формалізована постановка задачі

Необхідно розробити комплексну інтелектуальну систему, яка здатна класифікувати кандидата на його психологічну відповідність деякому виду професійної діяльності. Для визначення схильностей та прихованих якостей людини буде аналізуватися текст його резюме. При цьому його зміст буде порівнюватися з апріорно вірними даними, які будуть отримані під час аналізу ринку праці та занесені у відповідну базу даних.

Передбачається, що база даних буде зберігати інформацію про вимоги до вакансії у текстовому вигляді, тому для створення системи підтримки прийняття рішень будуть використанні алгоритми обробки природньої мови (NLP). Як класи розпізнавання обрані вакансії: менеджер, няня, бухгалтер.

Вибір класів розпізнавання обумовлений рівнем абстракції вимог до кожної з цих професій. Таким чином, вимоги до роботи няні є більш однозначними, ніж для роботи менеджера.

У загальному випадку послідовність дій СППР буде наступною:

1. На першому етапі, система підтримки прийняття рішень буде отримувати на вхід текст резюме кандидата.

1.1 Вхідний текст буде нормалізовано шляхом видалення стоп-слів, цифр, знаків пунктуації.

1.2 За допомогою методів обробки природньої мови з вхідного тексту буде виокремлено іменники, які, як припускаються, повинні описувати кандидата.

2. Попередньо оброблений текст буде порівнюватися з базою даних, за допомогою статистичних метрик.

2.1. Іменники з вхідного тексту будуть зіставлятися з записами з бази даних.

2.2. За допомогою статистичних метрик буде виявлено найбільш відповідний клас для вхідних даних.

3. На виході користувач отримає інформацію до якого класу найбільше схоже його резюме.

Таким чином користувач може отримати рекомендації при виборі місця роботи зарахунок автоматичного статистичного аналізу резюме та співставлення його з апріорно отриманими даними.

2 ОПИС МЕТОДУ ДОСЛІДЖЕННЯ

2.1 Формування та аналіз вхідних даних

Дослідження в сфері ефективного вибору персоналу з використанням інформаційних технологій потребують даних, які є специфічними для ринку праці у кожній країні. Таким чином не можливо порівнювати кандидатів, наприклад, з Індії та України. Із-за занадто різної культури роботи, пропозицій та вимог до робітників. Було прийнято рішення самостійно сформувати базу даних з відкритих джерел, яка б задовольняла основні вимоги поточного бакалаврського дослідження.

Перше, що варто обрати – це професії на які буде зроблений основний акцент. Річ у тім, що після аналізу ринку праці стало зрозуміло, що пропозицій багато, а їх назви дуже часто не є уніфікованими. Роботодавці дуже часто додають до назв посад додаткові, уточнюючі слова або словосполучення, наприклад, «дизайнер дитячих меблів». У свою чергу, це ускладнює пошук для потенційних кандидатів, адже за різними назвами можуть приховуватися одні і ті самі вимоги до майбутнього працівника.

Було прийнято рішення розглянути три категорії працівників:

- **Менеджери**. Достатньо обширний клас, який включає в себе спеціалістів різних сфер діяльності. У більшості випадках, основна робота менеджерів – це організація відділів та контроль над процесом роботи. Тим не менш, є велика різниця у специфіці вимог до менеджера в залежності від сфери його діяльності.
- **Няня**. Однозначний клас, де можливо формалізувати вимоги до працівника. Хоч вимоги і умови роботи можуть відрізнятися, але у загальному випадку можливо створити уніфікований психологічний портрет ідеальної няні.
- **Бухгалтер**. На перший погляд однозначний клас, де можливо формалізувати вимоги до кандидата, але на практиці від спеціалістів цієї

категорії можуть вимагати широкий спектр обов’язків, які враховують не тільки роботу з документами та цінними паперами, а й, наприклад, організацію персоналу. Із-за цього складно формалізувати психологічний портрет ідеального бухгалтера.

Варто помітити, що кожна з вибраних категорій має свій рівень абстракції, а отже по різному підходить до формалізації. Таким чином, «менеджер» - це максимально обширний клас, який може включати в себе велику кількість психологічних характеристик, які залежить від спеціалізації працівника. Наприклад, мінорні вимоги до менеджеру по персоналу та проектного менеджеру будуть відрізнятися. Проте головне для цієї професії – це уміння роботи з людьми, організація робочого процесу. Тобто передбачається, що основні психологічні риси будуть пов’язані з взаємодією з колективом. Для няні – це вміння забезпечити індивідуальні потреби дитини. Для бухгалтера – це вміння організації текстової та числової інформації.

Наступним кроком аналізу ринку праці було виявлення найкрупніших компаній України. Згідно [13] це: Київстар, Нова пошта, EPAM, Rozetka, COMFY, Samsung, GlobalLogic, Асбіс Україна, Укрпошта. У поточному дослідженні важливо, що кожна з цих компаній має вакансії на позиції бухгалтера та менеджера.

Останнім і найбільш трудомістким кроком формування бази даних буде збирання інформації з відкритих джерел про вимоги до працівників на обрані вакансії. Таким чином пропонується, розділяти вимоги для кандидата на технічні, які залежать від досвіду роботи та професійної компетенції працівника, та психологічні, які є частиною характеру людини. Основним критеріє психологічних вимог є їх слабоформалізованість, адже складно оцінити компетенцію працівника, наприклад, в комунікації та організації колективу. Для таких вимог критерію досвіду не достатньо, адже ефективність виконання таких дій напряму залежить від психологічного портрету кандидата.

Загальна мета посади: Ключова мета цієї посади полягає в розвитку каналу активного продажу та забезпеченні виконання планів продажу в цьому каналі. Кандидат повинен бути готовим приймати виклики та активно впроваджувати стратегії для досягнення поставлених цілей продажу.

Основні обов'язки та вимоги:

- Розробляти та впроваджувати стратегічний вектор розвитку каналу активного продажу.
- Ініціювати та впроваджувати партнерські і мотиваційні системи винагороди для партнерів.
- Заохочувати співпрацю та підтримувати партнерів для досягнення спільних цілей.
- Брати участь у BTL активностях для підтримки та зміцнення партнерських відносин.
- Мати не менше 3 років досвіду у продажу та не менше 2 років управлінського досвіду (люди/проекти/процеси).
- Демонструвати навички управління процесами, проектами та третіми сторонами, зокрема партнерами.
- Володіти високим рівнем комунікативних навичок та здатністю підтримувати партнерські відносини.
- Мати досвід управління проектами та розуміння особливостей продажу товарів та послуг в каналі активного продажу.
- Знати форми та методи проведення маркетингових кампаній, промоактивностей та заходів зі стимулювання збуту.
- Мати вміння користуватися офісними програмами MS Office.

Рисунок 2.1. Приклад виокремлення інформації з вакансії

Таким чином, на рисунку 2.1 показаний приклад вакансії на позицію менеджера в салоні Київстар. Зеленим кольором виділені вимоги до кандидата, які передбачають наявність певних рис характеру. Для менеджера – це уміння знаходити спільну мову з співробітниками. Синім кольором позначенні речення, які описують технічні вимоги до кандидата та його навичок.

Дані було зібрано з декілька сотень вакансій, що представлені на ринку праці. На рисунку 2.2. показана структура сформованої бази даних.

Рисунок 2.2. Структура бази даних

В таблиці 2.1 наведений опис усіх полів, які використані в базі даних.

Таблиця 2.1. Опис бази даних

Назва таблиці	Опис таблиці	Назва поля	Опис поля
organization	Зберігає інформацію про компанію	id	Ідентифікатор компанії
		name	Назва компанії
profession	Зберігає інформацію про професію	id	Ідентифікатор професії
		name	Назва професії
		id_org	Ідентифікатор організації, де присутня ця професія
vacancy	Зберігає інформацію про вакансію	id	Ідентифікатор вакансії
		id_prof	Наслідуємий ідентифікатор професії
		P-requirements	Речення, що описують вимоги до психологічного портрету кандидата
		T-requirements	Речення, що описують вимоги до технічних навичок кандидата

Таким чином, основною характеристикою психологічної та технічної відповідності кандидата до деякого роду діяльності стане текст його резюме, який буде порівнюватися з базою даних, для знаходження найбільш схожих іменних конструкцій.

2.2 Аналіз статистичних метрик при роботі з природньою мовою

Під час дослідження, як основну статистичну метрику, буде використано частотний аналіз документів (TF-IDF), який найчастіше використовується саме для виокремлення та порівняння текстових ознак. TF-IDF обчислює важливість кожного слова у документі щодо кількості його вживань у цьому документі та у всій колекції текстів. Де TF (Частота терміна) позначає, наскільки часто певне слово з'являється у цьому документі. Таким чином, TF вимірює важливість слова у контексті окремого документа. IDF (Зворотна частота документа) вимірює, наскільки унікальним є слово по всій колекції документів. Слова, які з'являються у більшості документів, мають низький IDF, оскільки вони не вносять великої інформаційної цінності. Цей метод дозволяє виділити найбільш інформативні слова та зрозуміти, які з них мають більшу вагу для свого тексту у контексті всієї колекції (2.1).

$$TF - IDF(t, d) = TF(t, d) * IDF(t), \text{де} \quad (2.1)$$

TF(t, d) - Частота терміна (TF) для слова "t" у документі "d"; IDF(t) – Зворотна частота документа (IDF) для слова "t".

Переваги використання TF-IDF для отримання ознак:

- Врахування інформативності слів: TF-IDF враховує як частоту слова у документі, так і його загальну рідкість по всій колекції. Таким чином, він допомагає виділяти ключові слова, які часто зустрічаються в цьому документі, але не надто поширені в інших.
- Усунення шуму: Слова, які зустрічаються у більшості документів (стоп-слова), мають низький IDF і, отже, низьку загальну вагу TF-IDF. Це дозволяє усунути шум і фокусуватися на важливіших словах.

Вибір методу отримання ознак залежить від конкретної задачі та характеристик текстових даних. У нашому випадку TF-IDF підходить добре для завдань, пов'язаних із вилученням ключових слів, кластеризацією та

класифікацією. Однак для складніших завдань, де важливі семантичні відносини між словами, варто розглянути використання більш сучасних методів, таких як Word2Vec. Ці методи, на відміну від TF-IDF, враховують семантичні зв'язки між словами та векторне уявлення слів.

Також вибір на користь TF-IDF обґрунтований тим фактом, що припускається поступове динамічне оновлення бази даних. У свою чергу це призведе до лінійного збільшення об’єму інформації, але в не значній мірі ускладнить обчислення для TF-IDF.

Не менш важливим аспектом розробки є вибір статистичної метрики для оцінки належності вхідного математичного опису до класу розпізнавання. У рамках NLP найчастіше використовують наступні оцінки:

Accuracy. Позначає частку випадків, коли модель робить правильний прогноз, порівняно із загальною кількістю прогнозів, які вона зробила. Найкраще використовувати, коли вихідна змінна є категорійною або дискретною. Наприклад, як часто алгоритм класифікації настроїв є правильним.

Precision. Оцінює відсоток істинних позитивних результатів, виявлених для всіх позитивних випадків. Особливо корисно, коли виявлення позитивних моментів важливіше, ніж загальна точність. Наприклад, при виявленні маловірогідних сценаріїв.

Recall. Відсоток справжніх позитивних результатів у порівнянні з комбінованими та хибно позитивними результатами. Корисно для вимірювання здатності моделі фіксувати всі відповідні позитивні випадки, але може бути чутливим до дисбалансу класів, віддаючи перевагу моделям, які передбачають усі випадки як позитивні.

F1 Score. Поєднує точність і запам’ятовування, щоб отримати єдиний показник — як повноту, так і точність. (2 * *Precision* * *Recall*) / (*Precision* * *Recall*). Використовується разом з точністю та корисний у завданнях позначення послідовності, таких як вилучення сутності та пошукові відповіді на запитання,

але чутливий до дисбалансу класів, на нього можуть впливати відносні ваги *Precision* та *Recall*

AUC. Поєднує істино позитивні та хибно позитивні результати, оскільки поріг для передбачення різний. Використовується для вимірювання якості моделі незалежно від порогу прогнозування та для знаходження оптимального порогу для завдання класифікації.

MRR. Середній взаємний ранг. Оцінює отримані відповіді з огляду на ймовірність їх правильності. Середнє обернене значення рангів отриманих результатів. Широко використовується в усіх завданнях пошуку інформації, включаючи пошук статей і пошук електронної комерції.

MAP. Середня точність, розрахована для кожного отриманого результату. Використовується в інформаційно-пошукових завданнях.

Word Error Rate (WER). Відношення кількості помилок у вихідних даних розпізнавання мовлення до загальної кількості слів у контрольній транскрипції. Широко використовується для оцінки моделей розпізнавання мовлення, але не чутливий до семантичних помилок.

Звісно, що на практиці не має потреби у використанні усіх зазначених статистичних метрик для оцінки результату класифікації, але на різних стадіях розробки кожна оцінка може бути по своєму ефективна. Враховуючи той факт, що у подальшому планується поступове збільшення об’єму бази даних.

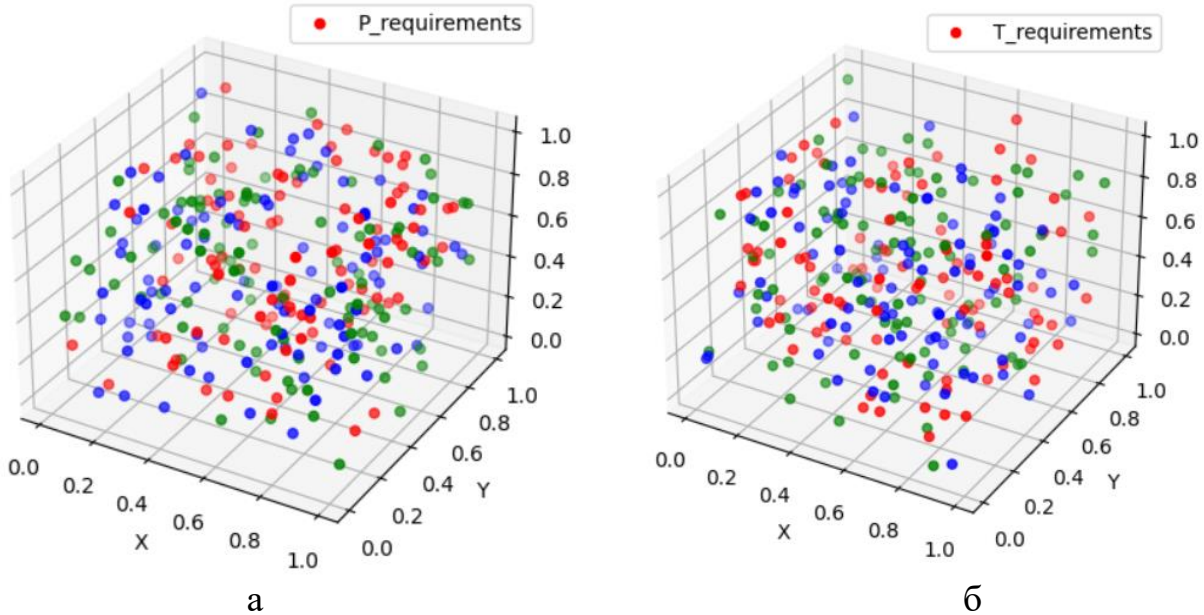
Отже для створення інтелектуальної компоненти буде використано алгоритм TF-IDF, який виокремить усі характерні слова-ознаки для кожного з класів розпізнавання, а для перевірки точності моделі буде застосовано *Accuracy, Precision, Recall* та *F1 Score.*

3 ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Підготовка вхідних даних

Для початку необхідно нормалізувати дані в таблиці 2.1. з полів P_requirements та T_requirements. Таким чином, щоб позбутися не інформативних слів та словосполучень. До яких відносяться сполучники, знаки пунктуації та числа.

Однак, враховуючи специфіку предметної області нас цікавлять лише іменники, бо тільки вони повинні давати характеристику кандидата. Тим не менш, серед них можуть траплятися іменники, які не характеризують кандидата. Наприклад, назви професій, компаній або ж методик. Для ідентифікації іменних сутностей було використано наступне API [14]. З його допомогою вдалося позбутися найменш інформативних категорій слів.

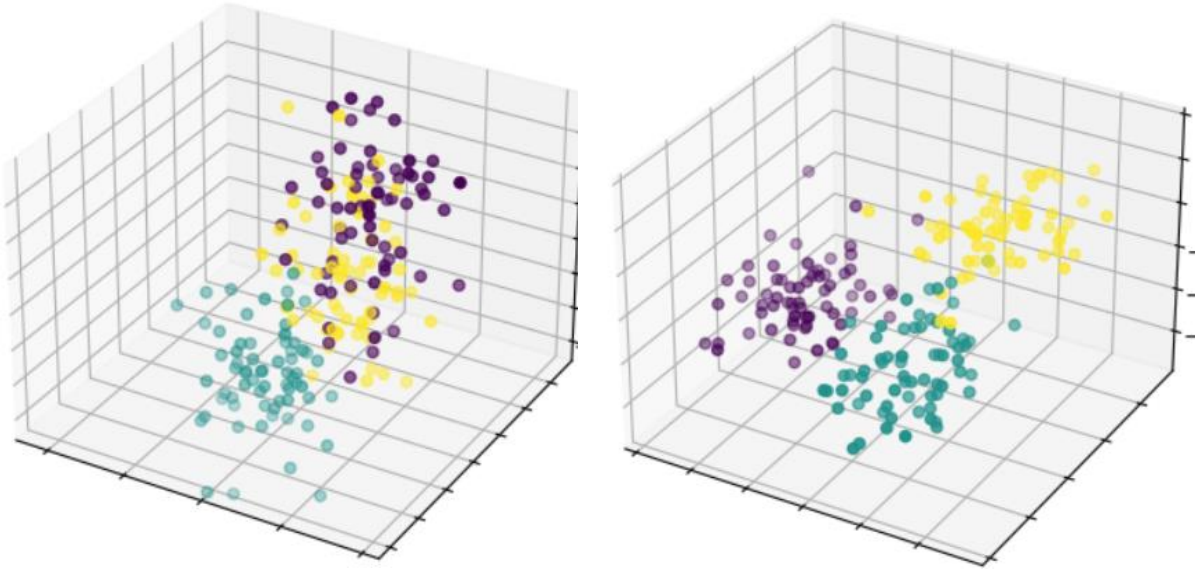


а б

Рисунок 3.1. Не нормалізовані вхідні дані: а) психологічні ознаки кандидата; б) технічні вимоги до кандидата

При аналізі рисунку 3.1 добре видно, що спочатку вхідні дані були погано структуровані. Це пов’язано з тим, що вони складалися з цілих речень, а отже містили значну кількість не інформативних ознак. У свою чергу, на рисунку 3.2.

показані ці самі дані, але вже після їх нормалізації, що призвело до структуризації та виокремлення чітких окремих класів розпізнавання.



а б

Рисунок 3.2. Нормалізовані вхідні дані: а) психологічні ознаки кандидата; б) технічні вимоги до кандидата

Аналіз рисунку 3.2 показує, що після нормалізації ознаки почали формувати чіткі центри. Це свідчить про те, що обрані класи розпізнавання вдалося розділити в просторі ознак, а отже є можливість побудувати точний класифікатор.

Додатково варто зауважити, що на рисунку 3.2.б, перетин класів розпізнавання в просторі ознак значно менший. Припускається, що це пов’язано з тим, що технічні вимоги більш точно характеризують вид діяльності, ніж психологічна характеристика. Наприклад, уміння підтримати розмову, є набагато більш багатозначним і частим вмінням, ніж володіння певною мовою програмування.

3.2 Програмна реалізація алгоритму класифікації для оцінки резюме

Для реалізації інтелектуальної компоненти СППР оцінки відповідності кандидатів використаємо частотний аналіз. Такий вибір обумовлений

припущенням, що більшість резюме є шаблонними, а отже будуть містити схожі синтаксичні конструкції. Насамперед, це стосується навичок людини, її характеру та вимог до роботи. У такому випадку прийняття рішень буде проводитися шляхом порівняння вхідних даних з емпірично отриманими даними, які відповідають кожному з наявних класів.

Перевірка функціональної ефективності системи для розпізнавання здійснювалася за допомогою частотного аналізу, де вхідний, нормалізований текст векторизувався і порівнювався з навчальною вибіркою. За допомогою статистичних метрик проводилася оцінка міри схожості вхідного математичного опису на кожен з класів розпізнавання. Отриманий результат навчання занесений в таблицю 3.1.

Таблиця 3.1 точнісні оцінки системи

клас	метрики				
	Accuracy	Precision	Recall	F1 Score	WER
няня	0.78	0.80	0.74	0.84	0.4
бухгалтер	0.81	0.73	0.8	0.76	0.64
менеджер	0.93	0.89	0.76	0.81	0.71

На вхід системі подамо наступний текст, який характеризує навички кандидата:

```
Професійні навички:
– володію комп’ютером (MS Office (Excel, Word),
інтернет);
– вміння користуватися офісною технікою;
– досвід роботи з клієнтами;
– досвід в сфері оптових і роздрібних продажів;
– навички Інтернет продажу;
– навички ведення документообігу;
– навички проведення рекламних компаній;
– грамотна письмова та усна мова;
– навички управління персоналом;
– вміння вільно вести переговори;
– Знання мов: українська – рідна; російська – вільно
володію; англійська – середній рівень, французька –
```

```
середній рівень.
Особисті якості:
Націленість на результат, уважність до деталей, вміння
працювати з великими обсягами інформації, ініціативність,
порядність.
```

Після етапу нормалізації вхідний текст прийме наступний вигляд:

```
навички ms office excel word інтернет користуватися офісною
технікою клієнтами оптових і роздрібних продажів Інтернет
продажу документообігу проведення рекламних компаній
грамотна письмова та усна мова управління персоналом вести
переговори українська російська англійська французька
націленість на результат уважність деталей вміння
працювати великими обсягами інформації ініціативність
порядність
```

Звісно, що текст частково втратив семантичний зміст, але у рамках поточної задачі це не настільки важливо, адже оцінка резюме проводиться за ключовими словами, а не за змістом. Така реалізація СППР обумовлена ідеєю, що найголовніші риси характеру, навички та вимоги до кандидата можливо сформулювати мінімальним переліком слів. Наприклад, через їх перерахування: ініціативність, мотивованість, програмування, веб-дизайн тощо. З цієї причини під час нормалізації ми залишаємо лише іменники, адже припускається, що інші частини мови, такі як прислівники, прикметники, дієслова, не є інформативними у резюме.

Наступним кроком нормалізації є використання алгоритму TF-IDF, який дозволить визначити найбільш цінні слова серед залишених іменних сутностей. Цього на попередньому етапі, бо інакше більшість з не інформативних частин мови були помилково класифіковані як інформативно значущі.

```
ms office excel word інтернет офісною клієнтами оптових
роздрібних Інтернет документообігу рекламних грамотна
письмова усна управління персоналом результат уважність
деталей вміння ініціативність порядність
```

Таким чином на етапі нормалізації вдалося зменшити початковий об’єм тексту майже в 4 рази, не втративши інформативність. Під час тестування

система віднесла вхідний текст до клас «менеджер» з вірогідністю в 0.71, що є правильним вибором.

Хоча система прийняття рішень показала цілком задовільний результат при тестуванні, але для підвищення вірогідності правильного прийняття класифікаційних рішень пропонується, по-перше, збільшити навчальну вибірку. По-друге, провести додаткове дослідження впливу нормалізації на ефективність класифікації, але вже з більшою кількістю даних.

Другим етапом покращення СППР є розширення алфавіту класів розпізнавання. Додавання нових вакансій та дослідження їх взаємозв’язку. Оскільки, припускається, що більшість професій мають схожі або ж повністю ідентичні вимоги до кандидатів.

Скоріш за все, у подальшому навчальну вибірку СППР потрібно буде структурувати. Таким чином, щоб поєднати схожі професії. Це б дозволило кандидату оцінити свої навички та отримати рекомендації щодо переліку можливих професій, де б він міг проявити себе максимально ефективно.

При збільшенні бази даних варто розглянути додаткову задачу вибору найбільш релевантної компанії для кандидата. Це можливо зробити через створення класифікаційних правил, які б містили лише психологічні характеристики, які прив’язані до організації.

ВИСНОВКИ

Було розроблено автоматизовану систему підтримки прийняття рішень для оцінки релевантності навичок та вмінь кандидата щодо вимог професії. Таким чином користувач може отримати рекомендації при виборі місця роботи зарахунок автоматичного статистичного аналізу резюме та співставлення його з апріорно отриманими даними.

Було проведено детальний аналіз предметної області. Огляд сучасних методологій оцінки психологічної придатності кандидата та співставлення їх з вимогами українського ринку праці.

Для реалізації системи підтримки прийняття рішень було створено унікальну базу даних, яка складається з характеристик вакансій для найбільших компаній України.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. [Електронний ресурс]. RESUME WORDED https://resumeworded.com/score
2. [Електронний ресурс]. ONLINE RESUME BUILDER. https://resume.io/
3. [Електронний ресурс]. Responsible AI that ensures your writing and reputation shinehttps://www.grammarly.com/?utm_source=google&utm_medium=cpc&utm_campaign=19835870348&utm_content=652228915101&utm_term=resume+review
4. [Електронний ресурс]. ВСЕУКРАЇНСЬКИЙ ПРОЄКТ З ПРОФОРІЄНТАЦІЇ ТА ПОБУДОВИ КАР'ЄРИ. https://hryoutest.in.ua/
5. [Електронний ресурс]. Let AI Choose Your Optimal Career Path from Your Resume. https://medium.com/@jainamshah142/let-ai-choose-your-optimal-career-path-from-your-resume-836f79568dd5
6. Chou, Yi-Chi, and Han-Yen Yu. "Based on the application of AI technology in resume analysis and job recommendation." 2020 IEEE International Conference on Computational Electromagnetics (ICCEM). IEEE, 2020.
7. [Електронний ресурс]. Google Maps API.https://github.com/googlemaps/google-maps-services-python
8. Heffner, Joseph, and Oriel FeldmanHall. "A probabilistic map of emotional experiences during competitive social interactions." Nature communications 13.1 (2022): 1718.
9. Bunderson, J. Stuart. "How work ideologies shape the psychological contracts of professional employees: Doctors' responses to perceived breach." *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 22.7 (2001): 717-741.
10. Савченко, Т. Р. Інформаційна технологія персоналізованого діагностування ранньої стадії раку передміхурової залози. MS thesis. Сумський державний університет, 2023.

11. Oblaqulovna, Ernazarova Gulnora, and Yuldasheva Inobat Ashuralievna. "Professional Psychological Portrait Of The Future Teacher." International Neurourology Journal 27.4 (2023): 208-217.
12. Shkurko, Tatyana A. "Socio-psychological analysis of controlling personality." Procedia-Social and Behavioral Sciences 86 (2013): 629-634.
13. [Електронний ресурс]. ТОП-10 українських компаній. https://ain.ua/2023/04/05/v-top-100-ukrayinskyh-kompanij-za-obigamy-v-2022-roczi-potrapyly-epam-rozetka-ta-inshi/
14. [Електронний ресурс]. WebAnno. https://webanno.github.io/webanno/releases/3.6.7/docs/user-guide.html
15. Dovbysh, A. S., et al. "Decision-Making support system for diagnosis of breast oncopathologies by histological images." Cybernetics and systems analysis 59.3 (2023): 493-502.
16. Naumenko, Igor, et al. "Information-Extreme Machine Learning of an On-board Ground Object Recognition System with a Choice of a Base Recognition Class." COLINS. 2022.
17. Chou, Yi-Chi, and Han-Yen Yu. "Based on the application of AI technology in resume analysis and job recommendation." 2020 IEEE International Conference on Computational Electromagnetics (ICCEM). IEEE, 2020.
18. Tran, Thanh Tung, et al. "Improving Human Resources' Efficiency with a Generative AI-Based Resume Analysis Solution." International Conference on Future Data and Security Engineering. Singapore: Springer Nature Singapore, 2023.
19. Dovbysh, Anatoliy, et al. "Information-extreme machine learning on-board recognition system of ground objects with the adaptation of the input mathematical description." CMIS. 2020.

ДОДАТОК

```
pip install git+https://github.com/Pangeamt/web_anno_tsv

TRAIN_DATA = []
ent_list = []
from web_anno_tsv import open_web_anno_tsv

tsv = '/content/vidvidav-ochakivskykh-kotykiv.tsv'

with open_web_anno_tsv(tsv) as f:
    for i, sentence in enumerate(f):
        #print(f"Sentence {i}:", sentence.text)
        ent_list_sen = []
        for j, annotation in
enumerate(sentence.annotations):
            ent_list_sen.append((annotation.start,annotat
ion.stop,annotation.label))
        ent_list.append(ent_list_sen)
        ent_dic = {}
        ent_dic['entities'] = ent_list[-1]

        TRAIN_DATA.append([sentence.text,ent_dic])

import spacy
from pathlib import Path
import random

model = None
model_dir=Path("model_ner")
n_iter=100

if model is not None:
    nlp = spacy.load(model)
    print("Loaded model '%s'" % model)
else:
    nlp = spacy.blank('uk')
    print("Created blank 'uk' model")

if 'ner' not in nlp.pipe_names:
    ner = nlp.create_pipe('ner')
    nlp.add_pipe(ner, last=True)
else:
    ner = nlp.get_pipe('ner')
```

```
for _, annotations in TRAIN_DATA:
    for ent in annotations.get('entities'):
        ner.add_label(ent[2])

other_pipes = [pipe for pipe in nlp.pipe_names if pipe != 
'ner']
with nlp.disable_pipes(*other_pipes):  # only train NER
    optimizer = nlp.begin_training()
    for itn in range(n_iter):
        random.shuffle(TRAIN_DATA)
        losses = {}
        for text, annotations in TRAIN_DATA:
            nlp.update(
                [text],
                [annotations],
                drop=0.5,
                sgd=optimizer,
                losses=losses)
        print(losses)

if model_dir is not None:
    model_dir = Path(model_dir)
    if not model_dir.exists():
        model_dir.mkdir()
    nlp.to_disk(model_dir)
    print("Saved model to", model_dir)

import pandas as pd
import re

data = pd.read_csv('dataset.csv')

sentenses = []

for sub in classes.text:
    string = re.sub('\(.*?\)', '', sub)
    sentenses.append(string)

data['NoSynonyms'] = sentences

import numpy as np
from tqdm.auto import tqdm, trange
import nltk
```

```
import string

def remove_punctuation(text):
    return "".join([ch if ch not in string.punctuation
else ' '
for ch in text])

def remove_numbers(text):
    return ''.join([i if not i.isdigit() else ' ' for i
in text])

import re

def remove_multiple_spaces(text):
    return re.sub(r'\s+', ' ', text, flags=re.I)

from nltk.stem import *
from nltk.corpus import stopwords
from pymystem3 import Mystem
from string import punctuation

stopwords = []
stopwords.extend(['…', '«', '»', '...'])

prep_text =
[remove_multiple_spaces(remove_numbers(remove_punctuation
(text.lower()))) for text in tqdm(data["NoSynonyms"])]

data['text_prepNoSynonyms'] = prep_text

data

from nltk.stem.snowball import SnowballStemmer
from nltk.tokenize import word_tokenize
nltk.download('punkt')
from nltk.tokenize import word_tokenize

def makeStemmed(texts):
    stemmed_texts_list = []
    for text in tqdm(texts):
        tokens = word_tokenize(text)
        stemmed_tokens = [stemmer.stem(token) for token
in tokens if token not in stopwords]
        text = " ".join(stemmed_tokens)
```

```
        stemmed_texts_list.append(text)
    return stemmed_texts_list

data['text_stem'] =
makeStemmed(data['text_prepNoSynonyms'])

X = data['text_stem'].values.astype('U')
y = data['mark']

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.4, random_state = 42)

from sklearn.pipeline import Pipeline

from sklearn.feature_extraction.text import
TfidfTransformer
from sklearn.feature_extraction.text import
CountVectorizer
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn.linear_model import LogisticRegression

logreg = Pipeline([('vect', CountVectorizer()),
                ('tfidf', TfidfTransformer()),
                ('clf', LogisticRegression(n_jobs=1,
C=1e5)),])

%%time
logreg.fit(X_train, y_train)
%%time
y_pred = logreg.predict(X_test)
print('accuracy %s' % accuracy_score(y_pred, y_test))
print(classification_report(y_test, y_pred))
from sklearn.feature_extraction.text import
TfidfVectorizer

from sklearn.decomposition import PCA
import plotly.express as px
```