Класифікація тексту

Misha Beherksy

3 червня 2017 р.

Анотація

This is abstract

1 Теоретичні основи е-е

На відміну від штучно створених мов, наприклад мов програмування чи математичних нотацій, мови, які ми використовуємо для спілкування, розвивалися з покоління в покоління, постійно видозмінюючись, а тому досить складно відслідкувати і встановити набір чітких конкретно визначених правил. Розробка алгоритмів, що дозволяють "розуміти"людські висловлювання дає змогу покращити велику кількість аспектів взаємодії людини та комп'ютера: передбачення вводу, розпізнавання тексту, пошук інформації в неструктурованому тексті, переклад з однієї мови на іншу, аналіз емоційного забарвлення тексту та багато іншого. Створюючи інтерфейси, що дозволяють людині більш ефективно використовувати комп'ютер, ми прискорюємо розвиток багатомовного інформаційного суспільства.

1.1 Методи класифікації даних

1.1.1 Проблема класифікації даних

Задача класифікації – формалізована задача, яка містить множину об'єктів (ситуацій), поділених певним чином на класи. Задана кінцева множина об'єктів для яких відомо, до яких класів вони відносяться. Ця множина називається вибіркою. До якого класу належать інші об'єкти невідомо. Необхідно побудувати такий алгоритм, який буде здатний класифікувати довільний об'єкт з вихідної множини. Класифікувати об'єкт – означає вказати номер (чи назву) класу, до якого відноситься даний об'єкт. Класифікація об'єкта – номер або найменування класу, що видається алгоритмом класифікації в результаті його застосування до даного конкретного об'єкту. В математичній статистиці задачі класифікації називаються також задачами дискретного аналізу. В машинному навчанні завдання класифікації вирішується, як правило, за допомогою методів штучної нейронної мережі при постановці експерименту у вигляді навчання з учителем (supervised machine learning). Існують також інші способи постановки експерименту – навчання без вчителя (unsupervised learning), але вони використовуються для вирішення іншого завдання – кластеризації або таксономії. У цих завданнях поділ об'єктів навчальної вибірки на класи не задається, і потрібно класифікувати об'єкти тільки на основі їх подібності. У деяких прикладних областях, і навіть у самій математичній статистиці, через близькість завдань часто не відрізняють завдання кластеризації від завдання класифікації.

Деякі алгоритми для вирішення задач класифікації комбінують навчання з учителем і навчання без вчителя, наприклад, одна з версій нейронних мереж Кохонена – мережі векторного квантування, яких навчають способом навчання з учителем.

Прогностичне моделювання – використання статистичних методів для передбачення деякого цільового значення. Зазвичай, мається на увазі передбачення деякої величини в майбутньому, хоча узагальнено це не грає жодної ролі і може бути застосовано до будь-якого типу невідомої події, незалежно від того, коли вона відбулася. В багатьох випадках задача зводиться до вибору найкращої моделі, що намагається здогадатися результат на основі набору вхідних даних, наприклад визначення того, чи є деякий лист електронної пошти спамом. Моделі можуть використовувати один чи декілька класифікаторів, щоб визначати приналежність даних до деякої множини. Сам термін прогностичної моделі широко перетинається з поняттями машинного навчання в наукових статтях та в контексті розробки програмного забезпечення. В промисловому середовищі даний термін швидше відноситься до поняття прогностичного аналізу.

1.1.2 Існуючі методи класифікації даних

В залежності від вхідних даних, для задач класифікації можна виділити такі категорії:

- Характеристичний опис – найпоширеніший випадок. Кожен об'єкт описується набором своїх характеристик, які називаються ознаками. Ознаки можуть бути числовими або нечисловими.
- Матриця відстаней між об'єктами. Кожен об'єкт описується відстанями до всіх інших об'єктів навчальної вибірки. З цим типом вхідних даних працюють деякі методи, зокрема, метод найближчих сусідів, метод потенційних функцій.
- Часовий ряд або сигнал є послідовність вимірів у часі. Кожен вимір може представлятися числом, вектором, а в загальному випадку – характеристичним описом досліджуваного об'єкта в цей момент часу.
- Зображення або відеоряд.

Зустрічаються і складніші випадки, коли вхідні дані представляються у вигляді графів, текстів, результатів запитів до бази даних, і т. д. Як правило, вони приводяться до першого або другого випадку шляхом попередньої обробки даних та вилучення характеристик. Щодо класифікації сигналів та зображень, то її також називають розпізнаванням образів.

В залежності від кількості класів, на які розбиваються вхідні дані, отримуємо такий поділ:

- Двокласова класифікація (бінарна класифікація). Найпростіший в технічному відношенні випадок, який служить основою для вирішення складніших завдань.

- Багатокласова класифікація. Коли число класів досягає багатьох тисяч (наприклад, при розпізнаванні ієрогліфів або злитого мовлення), завдання класифікації стає істотно важчим.
- Непересічні класи.
- Пересічні класи. Об'єкт може належати одночасно до декількох класів.
- Нечіткі класи. Потрібно визначати ступінь належності об'єкта кожному з класів, звичайно це дійсне число від 0 до 1.

Прикладом одного з методів, що використовуэться найчастіше, є наївний баєсівський метод (байєсівський класифікатор). Наївна байєсівська модель є ймовірнісним методом навчання. Імовірність того, що документ d потрапить у клас c записується як $P(c|d)$. Оскільки мета класифікації = знайти найбільш відповідний клас для даного документа, то в наївній байєсівській класифікації задання полягає в знаходженні найбільш ймовірного класу $c_m = \$c_m = \underset{c \in C}{argmax} P(c|d) \frac{P(d|c)Pc)}{P(d)} = \underset{c \in C}{argmax} P(d|c)P(c)$.

Обчислити значення цієї ймовірност ібезпосередньо немаожливо, оскільки для цього потрібно, щоб навчальна множина містила всі (або майже всі) можливі комбінації класів і документів. Однак, використовуючи формулу Байєса, можна переписати вираз для $P(c|d)$ у вигляд $c_m = \underset{c \in C}{argmax} P(c|d)$. Використовуючи навчальну множину, ймоварність $P(c)$ можна оцінити як $\hat{P}(c|d) = \frac{N_c}{N}$, тобто відношення кількості документів у класі до загальної кількості документів у навчальній множині. Але за допомогою навчальної множини можна лише оцінити ймовірність, але не знайти її точне значеннння.

1.1.3 Машинне навчання з учителем

Машинне навчання - узагальнена назва методів штучної генерації знань з досвіду. Штучна система навчається на прикладах і після закінчення фази навчання може узагальнювати. Тобто система не просто вивчає наведені приклади, а розпізнає певні закономірності в даних для навчання.

Серед багатьох програмних продуктів машинне навчання використовують: системи автоматичного діагностування, розпізнавання шахрайства з кредитними картками, аналіз ринку цінних паперів, класифікація ланцюжків ДНК, розпізнавання мовлення та тексту, автономні системи.

Машинне навчання — розділ штучного інтелекту, має за основу побудову та дослідження систем, які можуть самостійно навчатись з даних. Наприклад, система машинного навчання може бути натренована на електронних повідомленнях для розрізняння спам і не спам-повідомлень. Після навчання вона може бути використана для класифікації нових повідомлень електронної пошти на спам та не-спам папки.

В основі машинного навчання розглядаються уявлення та узагальнення. Представлення даних і функцій оцінки цих даних є частиною всіх систем машинного навчання, наприклад, у наведеному вище прикладі повідомлення по електронній пошті, ми можемо уявити лист як набір англійських слів, просто відмовившись від порядку слів. Існує широкий спектр завдань машинного навчання та успішних застосувань. Оптичне розпізнавання символів, в яких друковані символи розпізнаються автоматично, ґрунтуючись на попередніх прикладах, є класичним прикладом техніки машинного навчання. У 1959 році Артур Самуїл визначив машинне навчання як "Поле дослідження, яке дає комп'ютерам можливість навчатися, не будучи явно запрограмованим"Samuel [1959].

Практичне використання відбувається, переважно, за допомогою алгоритмів. Різноманітні алгоритми машинного навчання можна грубо поділити за такою схемою:

- Навчання з вчителем – алгоритм вивчає функцію на основі наданих пар вхідних та вихідних даних. При цьому, в процесі навчання, «вчитель» вказує вірні вихідні дані для кожного значення вхідних даних. Одним з розділів навчання з вчителем є машинна класифікація. Такі алгоритми застосовуються для розпізнавання текстів.
- Багатокласова класифікація. Коли число класів досягає багатьох тисяч (наприклад, при розпізнаванні ієрогліфів або злитого мовлення), завдання класифікації стає істотно важчим.
- Навчання без вчителя.
- Пересічні класи. Об'єкт може належати одночасно до декількох класів.
- Навчання з закріпленням (англ. reinforcement learning): алгоритм навчається за допомогою тактики нагороди та покарання для максимізації вигоди для агентів (систем до яких належить компонента, що навчається).

Узагальнення в цьому контексті є здатність алгоритму для виконання точно на нових, невідомих прикладах після тренування на навчальному наборі даних. Основна мета учня узагальнювати свій досвід.

Також існує поняття інтелектуального аналізу даних, що за своєю природою відрізняється від машинного навчання. Два терміни часто плутають, оскільки вони не рідко використовують ті ж методи і перекриття.

Вони можуть бути умовно визначені наступним чином: машинне навчання фокусується на прогноз, заснований на відомих властивостях, витягнутих з навчальних даних. Інтелектуальний аналіз даних (який є кроком виявлення знань у базах даних) фокусується на відкриття (раніше) невідомих властивостей даних.

Ці дві області перекриваються у багатьох відношеннях: інтелектуальний аналіз даних використовує безліч методів машинного навчання, але часто з дещо іншою метою. З іншого боку, машинне навчання також використовує методи інтелектуального аналізу такі як "неконтрольоване навчання"або як попередній крок оброблення для покращення точності навчальної системи. Велика частина плутанини відбувається з основних припущень: в машинному навчанні, виконання, як правило, оцінюється по відношенню до здатності відтворювати відомі знання, в той час як в інтелектуальному аналізі даних ключовим завданням є виявлення раніше невідомого знання. Необізнаний (неконтрольований) метод, який обчислюється по відношенню до відомих знань, буде легко перевершений керованими методами. В той час в типових ІАД завданнях, керовані методи не можуть бути використані через відсутність попередньої підготовки даних.

Деякі системи машинного навчання намагаються усунути необхідність в людській інтуїції під час аналізування даних, а інші обирають спільний підхід між людиною і машиною. Людська інтуїція не може бути повністю виключена, так як конструктору системи необхідно вказати, як дані повинні бути представлені і які механізми будуть використовуватися для пошуку характеристик даних.

Навчання з підкріпленням — це галузь машинного навчання натхненна біхевіористською психологією, що займається питанням про те, які дії мають виконувати програмні агенти в певному середовищі задля максимізації деякого уявлення про сукупну винагороду. Через свою загальність, дана проблема вивчається, вивчається багатьма іншими дисциплінами, такими як теорія ігор, теорія управління, дослідження операцій, теорія інформації, оптимізація на основі моделювання, багатоагентні системи, колективний інтелект, статистика та генетичні алгоритми. Галузь, що займається навчанням з підкріпленням, також називається наближеним динамічним програмуванням. Попри те, що проблема навчання з підкріпленням, вивчалась теорією оптимального управління, більшість досліджень стосувались саме існування оптимальних рішень та їх характеристики, а не навчання чи наближених аспектів. В економіці та теорії ігор, навчання з підкріпленням може використовуватись для пояснення того, як при обмеженій раціональності може виникати рівновага.

Навчання з учителем (англ. supervised learning) є одним із способів машинного навчання, в ході якого випробувана система примусово навчається за допомогою наявної множини прикладів «стимул-реакція» з метою визначення «реакції» для «стимулів», які не належать наявній множині прикладів.

Між входами та еталонними виходами (стимул-реакція) може існувати деяка залежність, але вона невідома. Відома лише кінцева сукупність прецедентів – пар «стимул-реакція», звана навчальною вибіркою. На основі цих даних потрібно відновити залежність (побудувати модель відносин стимул-реакція, придатних для прогнозування), тобто побудувати алгоритм, здатний для будь-якого об'єкта видати досить точну відповідь. Для вимірювання точності відповідей, так само як і в навчанні на прикладах, може вводитися функціонал якості.

Задача машинного навчання може бути представлена у вигляді експерименту. Даний експеримент являє собою окремий випадок кібернетичного експерименту зі зворотним зв'язком. Постановка даного експерименту припускає наявність експериментальної системи, методу навчання і методу випробування системи або вимірювання характеристик.

Експериментальна система у свою чергу складається з випробовуваної (використовуваної) системи, простору стимулів одержуваних із зовнішнього середовища та системи управління підкріпленням (регулятора внутрішніх параметрів). В якості системи управління підкріпленням може бути використано автоматичний пристрій, що регулюють (наприклад, термостат), або людину-оператора (вчитель), здатну реагувати на реакції випробовуваної системи і стимули зовнішнього середовища шляхом застосування особливих правил підкріплення, що змінюють стан пам'яті системи.

Розрізняють два варіанти: (1) коли реакція випробовуваної системи не змінює стан зовнішнього середовища, і (2) коли реакція системи змінює стимули зовнішнього середовища. На рис. 1 зображено загальний вигляд такої експериментальної системи.

В залежності від результуючих даних, отриманих від системи, можна виділити такі категорії класифікуючих систем:

- Множина можливих відповідей нескінчення (відповіді є дійсними числами або векторами). В даному випадку говорять про задачі регресії та апроксимації.
- Множина відповідей звичайна – задача класифікації та розпізнавання образів.
- Відповіді характеризують майбутню поведінку процесу або явища. В цьому випадку мова йде про задачі прогнозування (прогностичне моделювання).

Існують також вироджені системи, які характеризуються дещо зміненою поведінкою підкріплення інформації ("вчителя"):

- Система підкріплення з керуванням по реакції (R — керована система) — характеризується тим, що інформаційний канал від зовнішнього середовища до системи підкріплення не функціонує. Дана система, незважаючи на наявність системи управління, відноситься до спонтанного навчання, оскільки

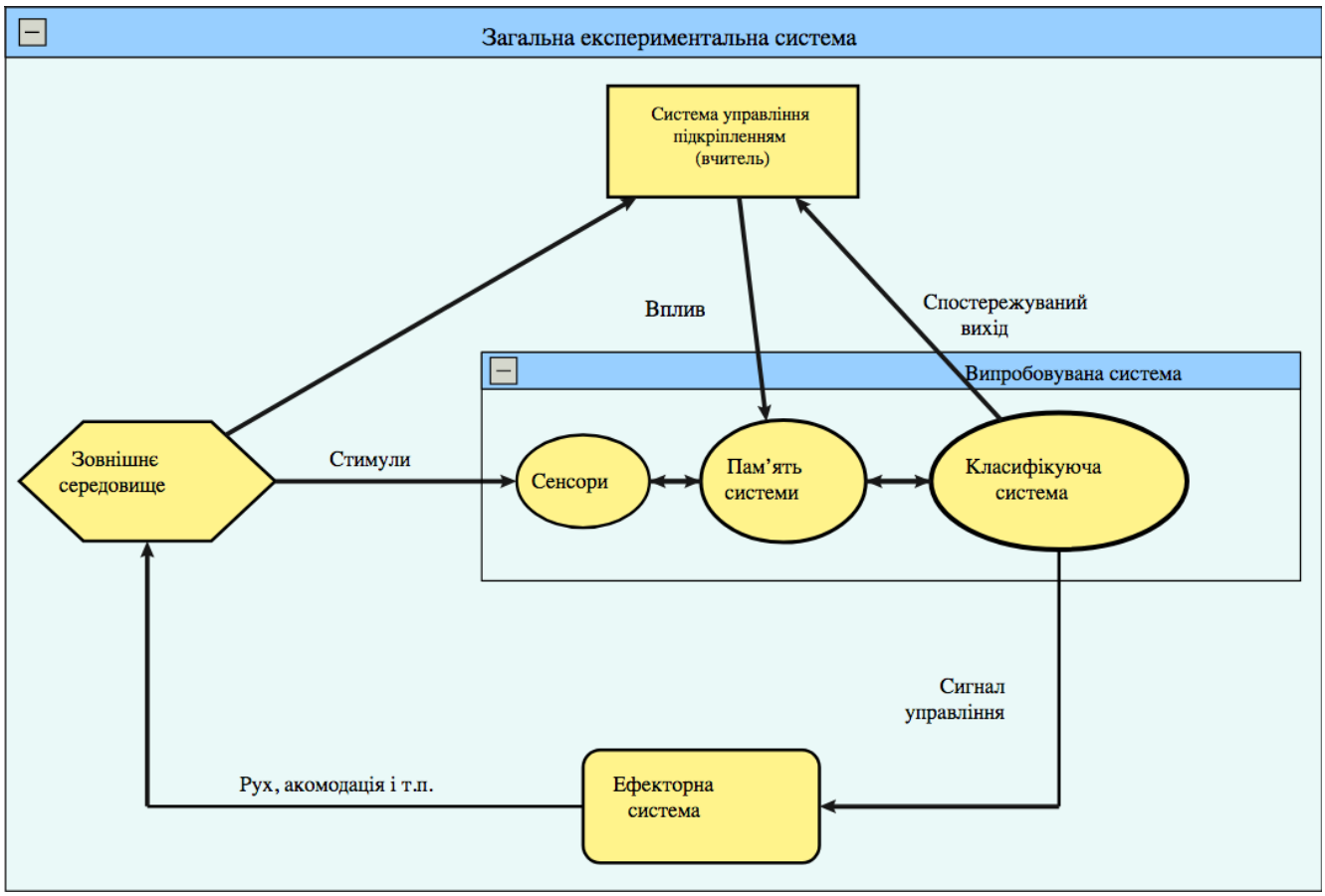

Рис. 1: Експериментальна система для навчання з учителем

випробовувана система навчається автономно, під дією лише своїх вихідних сигналів незалежно від їх "правильності". При такому методі навчання для управління зміною стану пам'яті не потрібно ніякої зовнішньої інформації.

- Система підкріплення з керуванням по стимулах (S — керована система) — характеризується тим, що інформаційний канал від випробовуваної системи до системи підкріплення не функціонує. Незважаючи на не функціонування каналу від виходів випробовуваної системи, відноситься до навчання з учителем, оскільки в цьому випадку система підкріплення (вчитель) змушує випробувану систему виробляти реакції згідно певного правила, хоча й не береться до уваги наявність істинних реакцій випробовуваної системи.

Дана відмінність дозволяє глибше поглянути на відмінності між різними способами навчання, оскільки грань між навчанням з учителем і навчанням без вчителя тонша. Крім цього, таке розходження дозволило показати для штучних нейронних мереж певні обмеження для S та R — керованих систем.

1.1.4 Класифікація текстів

Класифікація текстів (документів) – одне із завдань інформаційного пошуку, яке полягає в тому, щоб віднести документ до однієї чи декількох категорій на основі вмісту документу. Класифікація може здійснюватися повністю в ручному режимі або автоматично за допомогою створеного вручну набору правил, або ж за допомогою застосування методів машинного навчання. Варто відрізняти класифікацію текстів від кластеризації, в останньому випадку тексти теж групуються за деякими критеріями, але попередньо задані категорії відсутні.

Розглянемо згадані вище три основних підходи до задачі класифікації текстів.

По-перше, класифікація не завжди здійснюється за допомогою комп'ютера. Наприклад, у звичайній бібліотеці тематичні рубрики присвоюються книгам власноруч бібліотекарем. Подібна ручна класифікація дорога і непридатна у випадках, коли необхідно класифікувати велику кількість документів з високою швидкістю.

Інший підхід полягає в написанні правил, згідно яких можна зарахувати текст до тієї чи іншої категорії. Наприклад, одне з таких правил може виглядати наступним чином: “якщо текст містить слова похідна і рівняння, то віднести його до категорії математика”. Спеціаліст, який знайомий з предметною областю і володіє навичкою написання регулярних виразів, може скласти низку правил, які потім автоматично застосовуються до класифікації нових документів. Цей підхід краще попереднього, оскільки процес класифікації автоматизується і кількість оброблюваних документів стає практично не обмеженою. Більш того, побудова

правил власноруч може підвищити точність класифікації у порівнянні з машинним навчанням. Однак створення і підтримка правил в актуальному стані (наприклад, якщо для класифікації новин використовується ім'я чинного президента країни, то відповідне правило потрібно час від часу змінювати) вимагає постійного контролю зі сторони фахівця.

Нарешті, третій підхід ґрунтується на машинному навчанні. У цьому підході набір правил або, більш загально, критерій прийняття рішення текстового класифікатора обчислюється автоматично з навчальних даних (іншими словами, проводиться навчання класифікатора).

Навчальні дані – це деяка кількість наочних зразків документів з кожного класу. У машинному навчанні зберігається необхідність ручної розмітки (термін "розмітка" означає процес надання документу певного класу), але вона є більш простим завданням, ніж написання правил. Крім того, розмітка може бути проведена в звичайному режимі використання системи. Наприклад, у програмі електронної пошти може існувати можливість позначати листи як спам, таким чином формуючи навчальну множину для класифікатора – фільтра небажаних повідомлень. Тому класифікація текстів, заснована на машинному навчанні, є прикладом навчання з учителем, де в ролі вчителя виступає людина, що задає набір класів і розмічає навчальну множину.

Класифікація за змістом є класифікацією, в якій увага приділена конкретним питанням. У документі визначається клас, до якого його зараховують. Це, наприклад, правило бібліотечної класифікації: принаймні 20% від змісту книги має бути близько класу, до якого відноситься книга. В автоматичній класифікації – це може бути кількість разів, коли дані слова з'являються в документі.

Класифікація за запитом (або індексація) є класифікацією, в якій очікуваний запит від користувачів впливає на те, як документи класифікуються. Класифікатор запитує себе: "За якими дескрипторами цей об'єкт можна знайти?"Тоді оброблюються всі можливі запити та обираються найбільш відповідні. Поняття дескриптора в даному контексті означає лексичну одиницю (слово чи словосполучення) інформаційно-пошукової мови, яка служить для опису смислового змісту документів.

Класифікація за запитом може бути класифікацією, яка орієнтована на певну аудиторію або групу користувачів. Наприклад, бібліотека або база даних для феміністських досліджень можуть класифікувати (індексувати) документи по-різному в порівнянні з історичною бібліотекою. Це, ймовірно, краще, однак, класифікація робиться згідно деяких ідеалів і відображає мету бібліотеки або бази даних по класифікації. Таким чином, вона не обов'язково є видом класифікації або індексації на основі досліджень користувачів. Тільки якщо застосовуються емпіричні дані про використання чи користувачів, слід звернутися до орієнтованих класифікацій та розглядати в якості підходу користувача.

In text classification, we are given a description $\mathbb{X}$ of a document, where $\mathbb{X}$ is the document space ; and a fixed set of classes $\mathbb{C}$ Classes are also called categories or labels . Typically, the document space $\mathbb{X}$ is some type of high-dimensional space, and the classes are human defined for the needs of an application, as in the examples China and documents that talk about multicore computer chips above.

Тут реально матеріал з якоїсь книжки і реф на неї Doe [2100]

1.2 Exploratory data analysis

Візуалізація для наступних цілей: * Комунікативна - представлення даних та ідей - проінформувати - підтримати і аргументувати - вплинути і переконати * Дослідницька - вивчити (дослідити) дані - проаналізувати ситуацію - визначити наступні кроки - прийняти рішення стосовно деякого питання

$$\alpha = \sqrt{\beta} \tag{1}$$

1.3 Класифікація тексту

Метою класифікації текстів є розподіл документів на групи наперед визначених категорій. *-*

1.4 Учи матчасть

Результати показують, що стабільно показують чудові результати для завдань класифікації текстів, суттєво перевищуючи показники інших методів.

2 Розробка алгоритму

Оптимізація алгоритмів для використання у предметній галузі

3 Програмна реалізація

Особливості і деталі програмної реалізації

4 Аналіз рішення

Анализ, согласно критериям как работает, пути улучшения (таблица сравнения с существующими подходами, графики, диаграммы)

Зміст ідеї	Напрямки застовування	Вигоди для користувача
Система побудови універсальних пропонованим моделей та методів для підвищення їх ефективності та продуктивності роботи загалом	1. Використання науковцями та спеціалістами моделей та методів для подальших досліджень	Зручний та зрозумілий вихідний код нових алгоритмів, що дозволяє швидше та ефективніше, тим самим зосереджуючись на прикладних задачах, замість деталей реалізації
2. Узагальнення алгоритмів для роботи з різними типами даних	Універсальність моделі дозволить не перемикати контексти під час роботи з різними типами даних, використовуючи однаковий підхід для вхідної інформації	
3. Отримання кращих результатів передбачень для даних, що змінюються з часом	Допомога під час роботи з величинами, що залежать від часу: курси валют, показники біржі, зміни клімату	

Табл. 1: Опис ідеї стартап-проекту

№ п/п	Показники стану ринку (найменування)	Характеристика
1	Кількість головних гравців, од	1
2	Загальний обсяг продаж, грн/ум. од	914 218 млн грн
3	Динаміка ринку (якісна оцінка)	Спадає
4	Наявність обмежень для входу (вказати характер обмежень)	Висока доля невизначеності, відсутність попереднього досвіду та необхідних статистичних даних
5	Специфічні вимоги до стандартизації та сертифікації	-
6	Середня норма рентабельності в галузі (або по ринку), %	18-20%

Табл. 2: Попередня характеристика потенційного ринку стартап-проекту

5 Стартап

Опис ідеї стартап-проекту

numeric literals	integers	in decimal	8743
		in octal	0o7464
			0O103
		in hexadecimal	0x5A0FF
			0xE0F2
	fractionals	in decimal	140.58
			8.04e7
			0.347E+12
			5.47E-12
			47e22
char literals			'H'
			'\n'
			'\x65'
string literals			"bom dia"
			"ouro preto\nmg"

Ринок є доволі привабливим для входження: пристойна середня норма рентабельності, що трохи вищ аза середній банківський відсоток на вклади у гривні, а спадання ринку потенційно відкриває його для нестандартних інноваційних рішень, оскільки існує дуже висока необхідність в розробці універсального методу для відновлення зображень.

Обрано альтернативу 2 як таку, що має на увазі довше життя проекту.

В якості цільових груп обрано: 1 та 2.

5.0.1 Тут субсубсекція з висновками

Проведений детальний аналіз ринку та перспектив розвитку проекту дав змогу отримати такі результати:

- Існує можливість ринкової комерціалізації проекту, на ринку наявний попит на пропонований продукт.

№ п/п	Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
1	Необхідність для інвесторів знайти перспективний метод для вкладень	Люди, які мають фінансову можливість та зацікавленість робоити інвестиції у інноваційні проекти	Люди, які мають фінансову можливість та зацікавленість робити інвестиції у інноваційні проекти мають на меті збільшення свого капіталу, підвищення свого іміджу, а також долучитися до новітніх технологій, щоб бути у тренді	Необхідно розробити методику оцінювання та рекомендації, які б з високою ймовірністю розраховували потенційні необхідні інвестиції та шляхи попередження ключиових ризиків
2	Необхідність команди для побудови цього	Активні люди, які бажають втілити у життя свій проект	Необхідність проаналізувати всі ключові фактори, щоб визначити, чи доцільно реалізовувати проект та чи вдасться залучити спонсорів	Високоточний метод оцінки відновлення зображень, щоб визначити доцільність реалізації відновлення зображень

Табл. 3: Характеристика потенційних клієнтів стаптап-проекту

№ п/п	Фактор	Зміст загрози	Можлива реакція компанії
1	Попит	Не вдасться розробити унікальний метод, який би можна було застосовувати для будь-яких алгоритмів та адаптувати для роботи з різними типами даних	Розробка максимально універсального методу
2	Конкуренція	Можливість появи конкурентів з дуже схожими функціями, їх вихід на ринок раніше за нас	Доопрацювання якості розроблюваного методу з фокусом на зручність та простоту використання, розробка нових властивостей, яких немає у конкурента. Розгляд можливості об'єднання компаній для подальної спільної роботи.
3	Економічні	Зменшення доходу інвесторів, що призведе до зменшення кількості інвестицій	Моніторинг економічної ситуації у країні, пошук закордонних користувачів та адаптація для світового ринку

Табл. 4: Фактори загроз

№ п/п	Фактор	Зміст можливості	Можлива реакція компанії
1	Попит	Унікальність пропонованого функціоналу та додаткових можливостей при умові невисокої конкуренції дозволить захопити велику частку ринку, особливо зацікавивши додатком невеликих інвесторів (бізнес-ангелів) та команди проектів, які не потребують значних інвестицій	Адаптація до ринку, що розширяється, моніторинг новітніх розробок та ризиків, які тільки нещодавно з'явилися
2	Науково-технічні	Поява нових технологій, виникнення нових ринкових умов та факторів, які виявлять значний вплив на розвиток алгоритмів класифікації	Активне використання використання рішення; у випадку, якщо наше рішення буде адним з перших та матиме суттєві відмінності від аналогів, захист інтелектуальної власності розробників, патентування цієї технології та додання її до інтелектувальних активів проекту
3	Соціально-культурні	Велика популярність сфери роботи з даними та їх аналізу	Адаптація системи до розширення ринку, появи нових умов та технологій

Табл. 5: Фактори можливостей

- Ринок відкритий для інновацій, прослідковується позитивна динаміка ринку.
- Рентабельність роботи на ринку вища за прибутковість банківських вкладів, а отже приваблює як інвесторів, так і розробників для роботи над перспективним проектом.
- З огляду на потенційні групи клієнтів існує потенціал та перспектива входу на ринок.
- Істотні бар'єри для входження відсутні.
- В якості варіанту для впровадження для ринкової реалізації проекту доцільно обрати довгострокову роботу та утримання клієнтів, роботу над покращенням розробленого методу з використанням багатовимірного статистичного аналізу.
- Конкуренція практично відсутня, а конкурентноспроможність самого продукту достатньо висока. ...

Враховуючи описані вище ключові моменти, можна зробити висновок, що подальша імплементація даного проекту є доцільною та обґрунтованою.

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентноспроможною)
1. Тип конкуренції - чиста конкуренція	Велика кількість методів відновлення зображень, частина з яких є запатентованою інтелектувальною власністю	Звертати увагу на якість та універсальність методу відновлення зображень
2. За рівнем конкурентної боротьби - національний	Відновлення зображень не буде прив'язуватися до географічних показників	Акцент в рекламі на потреби жителів великих міст (столиці), таргетування на науковців та молодих дослідників, а також на високозабезпечених людей - потенційних інвесторів
3. За галузевою ознакою - внутрішньогалузева	Конкуренцію складають подібні методики розробки прогностичних моделей	Акцентувати увагу на незвичайність подачі послуг, а також зручність у використанні та надійність, яку вони забезпечують
4. Конкуренція за видами товарів - між бажаннями	Потенційні клієнти роблять вибір між звичними методами побудови моделей (яких дуже велика кількість) і відчувають складність у виборі найбільш доцільного методу	Чітко зрозуміти потреби та бажання кожної з груп цільової аудиторії та розробляти гнучку систему, яка задовольнятиме потреби всіх груп користувачів
5. За характером конкурентних переваг - нецінова	Акцент знаходиться на унікальності та якості послуг, що надаються, а також на перевагах, які отримує клієнт під час використання наших послуг	Робота над покращення методики побудови прогностичних моделей та підвищенням її універсальності
3. За інтенсивністю - не марочна	Продається втілення ідеї, а не певний бренд	Просування ідеї у соціальних мережах

Табл. 6: Ступеневий аналіз конкуренції на ринку

Складові аналізу	Прямі конкуренти в галузі	Потенційні конкуренти	Постачальники	Клієнти	Товари-замінники
	Прямих конкурентів немає, непрямі - різноманітні методи побудови прогностичних моделей	Нові розробки у галузі	Інвестори диктують умови розвитку ринку: ключова умова - проект повинен бути потрібним користувачам та приносити користь	Кількість зацікавлених клієнтів, рівень зацікавленості в такому типі послуг	Поява схожих дешевших або якісніших продуктів-конкурентів
Висновки	Прямих конкурентів немає	- можливості входу в ринок присутні, необхідно вирішити проблему пошуку та адаптації статистичних даних - необхідність розробки універсального методу, який може бути використаний як інвесторами, так і командою проекту	Успіх нашого проекту залежить від рівня довіри інвесторів та команд проекту до новітнього методу побудови прогрностичних моделей	Клієнти формують попит на таку послугу	Універсальних методів, які могли б замінити запропонований проект немає

Табл. 7: Аналіз конкуренції в галузі за М. Портером

№ п/п	Фактор конкурентноспроможності	Обгрунтування (наведення чинників, що роблять фактор для порівнянння конкурентних проектів значущим)
1	Фактор часу	Ідея є частково новою, для перейняття ідеї та втілення її у життя потенційним конкурентам знадобиться час
2	Фактор новизни товару	Початковий успіх продукту очікується через його новизну та інтерес цільової аудиторії до нових інноваційних рішень
3	Фактор якості послуг та надання інформації	Науковці та експерти з обробки даних потребують універсальний метод побудови прогностичних моделей

Табл. 8: Обґрунтування факторів конкурентноспроможності

№ п/п	Фактор конкурентноспроможності	Бали 1-20 Рейтинг товарів-конкурентів у порівнянні з іншими методами оцінювання	2	3	4	5	6
1	Фактор часу	15			+		
2	Фактор новизни товару	20		+			
3	Фактор якості послуг та надання інформації	17		+			

Табл. 9: Порівняльний аналіз сильних та слабких сторін методу

Сильні сторони: Якість послуг, що надаються Новизна послуг Можливість використання як інвесторами, і командою з розробки	Слабкі сторони: Відсутність статистичних даних та попереднього досвіду в реалізації подібних рішень
Можливості: Створення нової ринкової ніші Потреба у ефективному та компактному методі створення прогностичних моделей Необхідність закладати у бюджет можливі ризики та зміни ринкових умов	Загрози: Різка зміна ринку, поява нових стартапів, економічна криза

Табл. 10: SWOT-аналіз стартап-проекту

№ п/п	Альтернатива (орієнтовний комплекс заходів) ринкової поведінки	Ймовірність отримання ресурсів	Строки реалізації
1	- Ціль: отримання прибутку в короткостроковій перспективі - Конкуренція: цінова та партнерська (пропонуємо свої нові послуги розповсюдження інформації про партнерів - рекламні послуги) - Взаємодія з фірмами: активна боротьба за долю ринку, що належить конкурентам	В короткостроковому плані - велика В довгостроковому плані - значний ризик втратити долю ринку, якщо займатися лише ціновою конкуренцією	8-12 місяців після запуску проекту
2	- Ціль: захоплення частини ринку, підтримання її розміру та поступове нарощення об'ємів - Конкуренція: нецінова (акцент на тому, що пропонуємо інноваційні послуги) - Взаємодія з конкурентами: співпраця, активний моніторинг їх діяльності, при можливій появі реальних конкурентів можна запропонувати злиття компаній/проектів	Висока ймовірність отримання ресурсів та утримання їх протягом довгого проміжку часу. Більш ймовірний розвиток компанії та постійне покращення продукту	8-12 місяців після запуску проекту - для отримання перших фінансових надходжень від розповсюдження інформації про акції магазинів-партнерів, та їх реклама. Далі фінансові надходження прогнозовано регулярними

Табл. 11: Альтернативи ринкового впровадження стартап-проекту

№ п/п	Опис профілю цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит в межах цільової групи (сегменту)	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
1	Високозабезпечені люди, які зацікавлені у пошуку перспективних проектів для інвестування	Споживачі слідкують за найновітнішими технологіями, бажають бути в тренді та готові сприйняти новий продукт	Потенційно високий, інвестори хочуть бути впевненими у доцільності своїх інвестицій та подальшому отриманні прибутку	Практично відсутня	При наявності достойної та доручної реклами - досить просто
2	Ініціативні люди та науковці, які мають хорошу ідею в схожій сфері та хочуть втілити її у життя	Споживачі готові сприйняти продукт, так як зацікавлені у глибинному аналізі ситуації	Високий попит	Практично відсутня	При наявності достойної та доречної реклами - досить просто

Табл. 12: Вибір цільових груп потенційних споживачів

№ п/п	Обрана альтернатива розвитку проекту	Стратегія охоплення ринку	Ключові конкурентноспроможні позиції відповідно до обраної альтернативи	Базова стратегія розвитку
1	Захоплення, підтримання та захист частки ринку	Стратегія концентрованого маркетингу	- Новизна послуг - Доступність продукту - Простота в користуванні продуктом - Додаткові зручні аспекти, які враховуються під час розрахунку ефективності та інвестиційної привабливості побудови прогностичних моделей, що вигідно виділяють наш продукт серед конкурентів	Стратегія диференціації

Табл. 13: Визначення базової стратегії розвитку

№ п/п	Чи є проект "першопрохідцем"на ринку	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів	Чи буде компанія копіювати основні характеристики товару конкурента і які?	Стратегія конкурентної поведінки
1	Частково	Нові споживачі, частково забиратиме споживачів конкурентів	Частково. Новий метод оцінювання ефективності побудови прогностичних моделей буде агрегувати декілька методик аналізу, що дозволить оцінювати проекти більш точно з використанням більшої кількості факторів, що впливають на проект	Стратегія лідера

Табл. 14: Визначення базової стратегії конкурентної поведінки

№ п/п	Вимоги до товару цільової аудиторії	Базова стратегія розвитку	Ключові конкурентноспроможні позиції власного стартап-проекту	Вибір асоціацій, які мають сформувати комплексну позицію власного проекту (три ключових)
1	Універсальність методу оцінювання з точки зору інвестора	Стратегія диференціації	Врахування всіх аспектів оцінювання проекту з точки зору інвестиційної привабливості	- Ваші гроші ефективно працюють у інноваційному прогресивному проекті
2	Універсальність методу оцінювання з точки зору команди проекту	Стратегія диференціації	Врахування всіх аспектів оцінювання проекту з точки зору інвестиційної привабливості та життєздатності проекту, доцільність реалізовувати інноваційний проект	- Реальна можливість втілити у життя ідею завдяки глибинному аналізу ключових аспектів та пошуку інвесторів
3	Необхідність враховувати ризики проекту, ринкові та економічні умови, що швидко змінюються	Стратегія диференціації	Врахування ключових ризиків та ринкових умов завдяки розробленій системі коефіцієнтів	- Детальний облік ризиків та моніторинг ринкових умов дозволять уникнути передчасного закриття проекту

Табл. 15: Визначення стратегій позиціонування

№ п/п	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами (існуючі або такі, що потрібно створити)
1	Універсальний метод оцінювання ефективності та інвестиційної привабливості проекту, який буде корисний як для інвесторів, так і для команди проекту	Методика оцінювання дозволить уникнути передчасного закриття проекту та перевитрат люджету завдяки високоточній оцінці на ранніх етапах проекту	Оцінювання проекту як з точки зори витрат та ефективності їх використання командою стартапу, так і з урахуванням потенційних ризиків, прихованих стратегічних переваг на недоліків.

Табл. 16: Визначення ключових переваг концепції потенційного товару

№ п/п	Рівень цін на товари-замінники	Рівень цін на товари-аналоги	Рівень доходів цільової групи споживачів	Верхня та нижня межі встановлення ціни на товар/послугу
1	Безкоштовно	Безкоштовно	Більше 10000 грн/місяць	-

Табл. 17: Визначення меж встановлення ціни

№ п/п	Специфіка закупівельної поведінки цільових клієнтів	Функції збуту, які має виконувати постачальник товару	Глибина каналу збуту	Оптимальна система збуту
1	Купують право на використання методики	Зберігання, сортування, встановлення контакту інформування	Однорівневий	Залучена

Табл. 18: Формування системи збуту

№ п/п	Специфіка поведінки цільових клієнтів	Канали комунікацій, якими користуються цільові клієнти	Ключові позиції, обрані для позиціонування	Завдання рекламного повідомлення	Концепція рекламного звернення
1	Позитивне відношення до інновацій та швидкий розвиток технологій призводять до появи великої кількості нових методів, росту кількості даних і побудова прогностичних моделей стає більш актуальною	Соціальні мережі (facebook, twitter), тематичні ресурси	Універсальний метод побудови прогностичних моделей	Впевнити клієнта у тому, що метод є унікальним та універсальним	Повідомлення у соціальних мережах, статті на веб-ресурсах, короткі демонстраційні ролики

Табл. 19: Концепція маркетингових комунікацій

6 Висновки

Тут багато незрозумілих слів, ще більше води, ніж у всіх інших частинах диплому

Результати показують, що стабільно показують чудові результати для завдань класифікації текстів, суттєво перевищуючи показники інших методів.

Література

John Doe. The Book without Title. Dummy Publisher, 2100.

Robert W. Fairlie. Kauffman Index of Entrepreneurial Activity. Kansas City: Ewing Marion Kauffman Foundation, 2014.

Brad Feld. Startup communities: Building an entrepreneurial ecosystem in your city. Hoboken, NJ: John Wiley & Sons, 2012.

Young-Hoon Kwak. A brief history of Project Management. Greenwood Publishing Group, 2005.

Arthur Samuel. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, Volume 44, 1959.

Dane Stangler. The Economic Future just Happened. Kansas City: Ewing Marion Kauffman Foundation., 2009.

Martin Stevens. Project Management Pathways. Association for Project Management. APM Publishing Limited, 2002.

Шмидт С. Бирман Г. Капиталовложения. Экономический анализ ин-вестиционных проектов. М.: ЮНИТИ-ДАНА, 2003.

Лапыгин Ю. Н. Управление проектами: от планирования до оценки эффективности. М.: Омега-Л, 2008.