МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

Економічний факультет

Кафедра статистики, інформаційно-аналітичних систем і демографії

КВАЛІФІКАЦІЙНА МАГІСТЕРСЬКА РОБОТА

СТАТИСТИЧНА КЛАСИФІКАЦІЯ КОМПАНІЙ БІОТЕХНОЛОГІЧНОГО НАПРЯМКУ ЗАСОБАМИ МАШИННОГО НАВЧАННЯ

Студента ІІ курсу
спеціальності 051 Економіка
за освітньо-науковою програмою
«Економічна аналітика та статистика»
денної форми навчання
Юрія ГОРОБЦЯ

(підпис)

Науковий керівник:
к.е.н., доцент.
Ігор ГОНЧАР

(підпис)

Роботу допущено до захисту на засіданні ЕК рішенням кафедри статистики, інформаційно-аналітичних систем і демографії, протокол № 12 від «12» травня 2022 року

Завідувач кафедри д.е.н., проф. Наталія КОВТУН

Київ – 2022

РЕФЕРАТ

Кваліфікаційна робота магістра містить 36 с., 3 рис., 6 табл., 16 джерел.

Ключові слова: машинне навчання, обробки природної мови, класифікація тексту, zero-shot classification.

Предмет дослідження: підходи до класифікації текстів з використанням методів машинного навчання та з використанням наперед натренованих zero-shot classification моделей.

Об'єкт дослідження: тексти зі змістом біотехнологічного напрямку.

Мета дослідження: оцінка доцільності використання попередньо натренованих моделей обробки природної мови для класифікації текстів біотехнологічного напрямку.

Методи дослідження базуються на використанні методів машинного навчання зі вчителем.

Наукова новизна: досліджено використання попередньо натренованих моделей обробки природної мови загального призначення для класифікації текстів біотехнологічного напрямку та порівняно із використання таких моделей із традиційними методами машинного навчання зі вчителем.

Практична цінність: розробка та виявлення моделей обробки природної мови, що можуть забезпечити якісну класифікацію текстів біотехнологічного напрямку з огляду на вимоги щодо влучності та чутливості класифікації.

RESUME

Taras Shevchenko National University of Kyiv
Faculty of Economics, Department of Statistics and Demography

Keywords: machine learning, natural language processing, text classification, zero-shot classification.

Master's qualification work «Statistical rating of biotechnological companies with machine learning tools» consists of 36 pages, 3 figures, 6 tables, 16 sources.

The subject of the research is text classification techniques using both supervised machine learning and zero-shot classification models. The object of the research is text with biotechnological contents.

The purpose of the research is to analyze general NLP zero-shot classification models usage feasibility for biotechnological text classification by comparing those models with traditional supervised machine learning techniques. The author has developed custom machine learning models for text classification using naïve Bayes classifier, logistic regression and neural networks and evaluated precision and recall of existing pretrained zero-shot classification models for texts with biotechnological contents.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

NLP – Natural Language Processing, обробка природної мови

НКБ – Наївний класифікатор Байєса

ReLU – Rectified Linear Units

ZSC – zero-shot classification

ВСТУП

Задача класифікації об'єктів та явищ супроводжувала людство з початку його існування. Проте з середини минулого сторіччя кількість даних почала стрімко збільшуватися і навіть набувати характеру експоненційої кривої. Це призвело до того, що сьогодні жодна людина фізично не може їх обробити, а тому з'явилися методи комп'ютерної обробки. Це стосується і класифікації текстів.

Традиційний підхід до класифікації текстів полягає у побудові власних моделей машинного навчання. Проте побудова власних моделей супроводжується необхідністю витрат на апаратне забезпечення, володіння навичками побудови моделей, а при побудові моделей з учителем додається необхідність мати значення залежної змінної для тренувальної та тестової вибірок (промарковані дані). Альтернативним підходом до класифікації текстів є використання вже натренованих моделей обробки природньої мови, зокрема моделі виду zero-shot classification. Їх використання дозволяє зменшити витрати на етапі тренування та витрати на збір промаркованих даних.

Тема використання zero-shot classification моделей досліджувалася такими вченими як Chang (2008) [1], Srivastava (2018) [2], Levy (2017), Obamuyide (2018) [3], Yin (2019) [4], командою Hugging Face [5], [6] та ін. В той же час, моделі як правило будувалися на основі корпусів текстів із загальновживаною лексикою. Цей факт, в свою чергу, ставить під сумнів можливість використання попередньо натренованих моделей для корпусів текстів із специфічними термінами, як-то тексти біотехнологічного напряму, через потенційну невідповідність лексикону, що був використаний на етапі тренування, та лексикону кінцевого використання моделі. Тому порівняння якості наперед натренованої моделі, яка дозволяє користуватися нею без збору промаркованих даних, з власними моделями, що тренувалися на таких даних, може відкрити широкі межі для спрощення класифікації у бізнес процесах.

Метою роботи була оцінка доцільності використання попередньо натренованих моделей обробки природної мови для класифікації текстів біотехнологічного напрямку.

Для досягнення цієї мети були поставлені такі задачі:

1. Зібрати корпуси текстів-описів компаній біотехнологічного напрямку із промаркованими цільовими бінарними класами.
2. Підготувати бази даних з використанням підходів обробки природної мови до класифікації із застосуванням моделей машинного навчання.
3. Натренувати моделі класифікації текстів із використанням методів логістичної регресії, наївного класифікатора Байєса та нейронної мережі та оцінити їх точність.
4. Оцінити якість класифікації за допомогою наперед натренованих моделей zero-shot classification та порівняти її з якістю власних моделей.

Об'єкт дослідження: тексти із змістом біотехнологічного напрямку.

Предмет дослідження: підходи до класифікації текстів з використанням методів машинного навчання та з використанням наперед натренованих zero-shot classification моделей.

Інформаційну базу кваліфікаційної магістерської роботи склали дані з текстами-описами діяльності компаній, що знаходяться у відкритому доступі бази даних Crunchbase [7].

Робота складається зі вступу, трьох розділів, висновків, списку використаних джерел. Перший розділ присвячено проблемі класифікації, зокрема класифікації компаній за напрямком її діяльності, огляду біотехнологічного сектору та сектору охорони здоров'я на ринку, теоретичним засадам підходів обробки природної мови та основних методів класифікації з використанням технологій машинного навчання.

У другому розділі обґрунтовується вибір методів класифікації та описується їх використання, процес збору та підготовки бази даних, методи оцінки точності класифікаційних моделей, а також використання наперед натренованої ZSC моделей.

В третьому розділі представлені результати класифікації текстів шляхом моделювання власних моделей та шляхом використання попередньо натренованих моделей.

РОЗДІЛ 1. ТЕОРЕТИКО-МЕТОДОЛОГІЧНІ ЗАСАДИ МЕТОДІВ ОБРОБКИ ПРИРОДНОЇ МОВИ

1.1. Суть та способи класифікації

Від початку свого існування людина тяжіє до абстрагування навколишніх явищ. Для того, щоб краще зрозуміти навколишній світ і мати змогу швидко на нього реагувати, вона намагається його помістити у певну систему: окремо живі та неживі істоти, окремо тверді речовини, рідини і гази, люди і тварини, знайомі і чужинці, день і ніч і т. д. Такий умовний поділ завжди заснований на певній ознаці (або наборі ознак) та дозволяє швидко реагувати на подібні об’єкти. З часом потреба такого розмежування тільки зростала і воно ставало більш складним. Сьогодні нам важливо знати надійний чи ненадійний позичальник, шахрайська чи законна транзакція, позитивний чи негативний відгук тощо. Всі ці приклади можна вважати класифікацією об’єктів, а сам процес розділення їх на групи – вирішенням класифікаційної задачі.

Задача класифікації – це задача розбиття множини об'єктів або явищ на апріорно задані групи, що називаються класами, всередині кожної з яких вони вважаються схожими один на одного, та мають приблизно однакові властивості й ознаки. Для класифікації наявність наперед визначених класів є обов’язковою. Класифікація буває бінарною (коли існує тільки два класи) та множинною (коли існує більше двох класів), причому множинну класифікацію часто можна спростити до бінарної.

Через постійне зростання кількості даних, які потрібно опрацьовувати, сучасна людина фізично не може це зробити. Тому вже винайдено велику кількість методів комп’ютерної обробки даних і багато ще належить винайти. Це твердження також стосується задачі класифікації. Саме через це виник напрямок машинного навчання. Для її вирішення можна використовувати різні методи, зокрема логістичну регресію, наївний класифікатор Байєса, дерево рішень, штучні нейронні мережі тощо.

Для вирішення таких завдань можна застосувати два підходи машинного навчання: з учителем і без учителя.

У першому випадку модель вчиться на попередньо розмічених даних. Тобто це означає, що має бути тренувальна множина та тестова множина з відомими залежними та незалежними змінними для того, щоб «натренувати» модель, визначати потрібні класи, а також перевірити чи правильно вона «навчилася». Для застосування такого підходу обов’язковою початковою умовою є наявність великої кількості даних з відомими класами.

У другому випадку алгоритм самостійно буде знаходити схожість між об’єктами та об’єднувати їх у певні групи. Недолік цього підходу в тому, що заздалегідь точно невідомо, за яким принципом модель вирішить розділити об’єкти (на які саме класи і чи на ті, які потрібні).

Незалежно від обраного підходу, на класифікаційну модель також буде впливати специфіка текстів, що обробляються. В наступному підрозділі розглянемо особливості біотехнологічних компаній та розглянемо приклади лексикону таких текстів.

1.2. Особливості класифікації текстів біотехнологічного напрямку

Сектор охорони здоров’я включає компанії, які виробляють медичні товари, такі як ліки та медичні прилади, і компанії, які надають медичні послуги: медичне страхування або лікування. У наш час сектор охорони здоров'я є одним з найбільших гравців на ринку: його капіталізація на момент початку 2020 року склала $6.26 трлн та поступилася лише сектору інформаційних технологій та фінансів ($9.58 трлн та $7.53 трлн відповідно) [8]. На сектор охорони здоров'я припадає понад 10% ВВП більшості розвинених країн, у Сполучених Штатах Америки він вважається найбільшим роботодавцем і на нього припадає 18% ВВП [9]. Отже, сьогодні охорона здоров'я вважається однією з найбільших та швидкозростаючих індустрій. Крім цього, з моральної точки зору, її ріст і розвиток є

дуже важливим для покращення рівня життя світового населення. Все це робить охорону здоров'я дуже привабливою для інвестування.

У секторі охорони здоров'я, як і у будь-якому іншому, існує значна потреба у класифікації. Наприклад, WHO (World Health Organization, Всесвітня організація охорони здоров'я) має окремі стандартизовані класифікаційні системи: ICD (International Statistical Classification of Diseases and Related Health Problems, Міжнародна статистична класифікація хвороб і пов'язаних з ними проблем зі здоров'ям), ICHI (International Classification of Health Interventions, Міжнародна класифікація медичних втручань) та інші [10].

Оскільки у світі існує дуже багато різноспрямованих компаній, постає проблема їх автоматичної класифікації. Існують компанії, інвестиційні фонди, які працюють тільки або переважно з охороною здоров'я. Для них важливо розуміти ситуацію на ринку, а тому знати які саме представники на ньому існують. Для цього потрібно виокремити саме ті компанії, які їх цікавлять, і зробити це потрібно автоматично, адже людина фізично не зможе сформувати повний список.

Проте для такої класифікації потрібно мати велику базу даних вже розмічених компаній, збирати яку може виявитися довгим та економічно затратним процесом, тому у цій роботі було вирішено порівняти точність різних підходів до класифікації корпусів текстів.

Розглянемо наступний приклад тексту-опису компанії, що виробляє продукти в стоматологічній галузі: «Revolutionizing healthcare services for insurance and medical companies. Telemedico is one of the largest providers of telemedicine in Europe. We deliver comprehensive telemedicine solutions for insurers, assistance companies, employers, medical entities, pharmaceutical companies, and as part of national healthcare systems. Taking into account the expectations of our business customers and the specificities of local markets, we offer innovative telemedicine solutions, their implementation, and high-quality services of remote doctor consultations. Our solutions are used by an ever-growing number of patients across the world». Це опис компанії-провайдера послуг в сфері дистанційної медичної допомоги.

Іншим прикладом тексту є: «Miambe features a minimally invasive antral membrane balloon elevation. The MIAMBE technique has been used by thousands of practitioners and extensively researched for over a decade making sinus lift an easier and safer procedure. With its system the absence of sinus floor and cases of 0-4 mm of residual bone under the antral membrane can be resolved with excellent results. The whole procedure is conducted through a 3mm diameter osteotomy. The MIAMBE balloon is the only balloon for dental use which is made of Silicon as used in cardiology». Текст містить професійні словосполучення, як-то «antral membrane», «sinus floor», «residual bone». Водночас, окремі слова «floor», «residuals», «bone» можуть вживатися в повсякденній мові. Тобто, моделі машинного навчання, побудовані на текстах загальновживаної лексики можуть мати труднощі із інтерпретацією такого роду текстів.

Для застосування методів машинного навчання до будь-якого корпусу текстів, чи із загальновживаними словами, чи із специфічними термінами, потрібно спершу провести первинну обробку тексту. Така обробка дозволить стандартизувати вільну людську мову та зробить можливим застосування до неї математичного апарату. В наступному підрозділі буде розглянуто розповсюджені методи обробки природньої мови.

1.3. Методи обробки природньої мови

Обробка природньої мови (NLP) – загальний напрямок штучного інтелекту та математичної лінгвістики, що вивчає комп’ютерне розпізнавання, аналіз та синтез природньої мови, а також можливість генерування текстів, у тому числі і машинний переклад.

Сьогодні галузь NLP – це перспективний напрямок, який ставить багато нових викликів для дослідників, оскільки тексти часто містять у собі певні частини, які не можна сприймати у прямому сенсі (наприклад, фразеологізми), які можуть містити непрямий порядок слів, займенники і т. п. Загалом щоб зрозуміти мову, часто потрібно мати широке розуміння навколишнього світу.

На поточному етапі розвитку технологій обробки природньої мови, використання моделей машинного навчання потребує попередньої обробки тексту, перетворення вільного тексту до такого формату, що може сприйматися математичним апаратом

Першим етапом такої обробки є переведення всього тексту до єдиного регістру, зазвичай до нижнього, щоб слова в різних регістрах, як-то «Компанія» та «компанія» однаковим чином впливали на класифікатори.

Другим етапом є перетворення вільного тексту до масиву окремих слів. Такий процес називається токенізацією (tokenization). Під час токенізації речення розбивається на окремі частини знаками пунктуації: пробілами, крапками, комами, та ін. В результаті утворюється масив токенів, до якого можуть входити не лише слова, а й числа, знаки пунктуації та інші частини речення.

Наступним етапом зазвичай є лематизація – уніфікація морфологічних форм слів: перетворення граматичної множини в однину (companies => company), або перетворення відмінюваних форм дієслів в початкову (works, worked => work). Для проведення якісної лематизації потрібні спеціальні словники.

Після цього доцільним кроком може бути видалення часто вживаних слів, які не несуть додаткового смислового навантаження. Наприклад, в англійській мові такими є артиклі, сполучники, займенники, допоміжні слова, базові дієслова. Перевагою видалення стоп слів є зменшення розмірності словника; видалення зайвих предикторів, що не повинні впливати на класифікатори. Видалення стоп-слів є необов’язковим.

Після первинної обробки, текст став більш стандартизованим, проте ми все ще маємо справу із представленням інформації у вигляді слів. Для застосування алгоритмів машинного навчання наступним необхідним кроком є перетворення даних із текстового у чисельне представлення. Існує декілька способів такого перетворення, у роботі буде розглянуто підхід «мішку слів» (bag of words) та підхід векторного представлення слів (word embeddings).

Bag of words полягає у підрахунку кількості різних слів в одному текстовому документі, без збереження зв’язку між словами та без збереження порядку слів.

Текст перетворюється на матрицю чисел, де рядок відповідає окремому текстовому документу, а стовпець — окремому унікальному слову. На перетині рядка і стовпця маємо кількість входжень слова у відповідному документі.

Ідея використання bag of words полягає в припущенні, що документи є змістовно схожими, якщо вони мають схожий набір слів всередині, та що за набором слів ми маємо змогу дізнатися про семантичне наповнення документа. В нашому дослідженні, із нашим корпусом текстових документів таке припущення відповідає дійсності, тому ми маємо можливість використання bag of words в нашому дослідженні.

Підхід bag of words має низку недоліків, як-то відмову від порядку слів, отже ігнорування контексту, а отже і ігнорування значення слів у документі. Зокрема, семантично схожі слова, наприклад, «король» та «королева», за цим підходом мають абсолютно незалежний вплив на модель машинного навчання. Іншим недоліком є залежність від кінцевого набору слів із тренувальної множини. У роботі bag of words був використаний під час будування моделей із використанням наївного класифікатору Байєса та логістичної регресії.

Альтернативним підходом до використання підходу bag of words є використання векторного представлення слів (word embeddings), що полягає у представленні корпусу документів у вигляді векторів, що належать до одного, векторного простору спільного для всього корпусу документів. Векторна модель використовується для вирішення багатьох завдань інформаційного пошуку, як-то пошук документа за певним запитом, вирішення задач класифікації та кластеризації документів.

Слова-вектори є представленням слів у вигляді чисел, із збереженням семантичного зв'язку між словами. Векторне представлення ґрунтується на близькості контекстів: якщо слова зустрічаються в тексті поруч з однаковими словами (тобто мають схожий контекст), то такі слова будуть мати близькі вектори. Наприклад, для вектору слова «король» одним з найближчих буде вектор слова «королева». Однак вектор слова «компанія» буде сильно відрізнятися від вектору слова «король». Більш того, різниця між векторами слів «король» та «королева»

буде схоже на різницю векторів слів «чоловік» та «жінка». Ця схожість та відмінність обумовлена частотою використання слів в різних контекстах. Така модель представлення слів у вигляді векторів була представлена в 2013 році Томашом Міколовим [11]. У роботі було використано word embeddings для будування власних штучних нейронних мереж.

В наступних розділах буде розглянуто, які саме методи машинного навчання використовувалися для будування власних моделей, та буде розглянуто існуючі вже натреновані моделі обробки та розуміння природної мови.

РОЗДІЛ 2. МЕТОДОЛОГІЧНІ ЗАСАДИ ВИКОРИСТАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ ОБРОБКИ ПРИРОДНОЇ МОВИ

2.1. Основні методи класифікації текстів

Як було зазначено вище, існує два підходи до класифікації: навчання з учителем і без учителя. Навчання з учителем зазвичай демонструє високі показники точності, проте той факт, що для нього потрібні промарковані дані, часто ускладнює його використання. Навчання ж без учителя не має такого недоліку. Принцип його роботи лежить у тому, що алгоритм самостійно знаходить певні патерни схожості у вхідних даних та розділяє дані на певні схожі групи. Проте навчання без учителя також є свої недоліки. Основною проблемою такого підходу є складнощі в оцінці точності; при спробі оцінити точність, вона здебільшого може виявитися незадовільною, і настане потреба у людській обробці результатів. Тому в даній роботі власні моделі без учителя не будувалися.

Для обробки природньої мови серед методів навчання з учителем найбільш вживаними є алгоритми глибинного машинного навчання, а також наївний класифікатор Байєса (НКБ) та логістична регресія.

Моделі побудовані із глибинним машинним навчанням здатні вирішувати широкий спектр задач: переклад між мовами, передбачення наступних слів, та, звісно, класифікацію текстів. Алгоритми глибинного машинного навчання демонструють найкращі результати у порівнянні із іншими методами, проте потребують більш великих обсягів тренувальних даних та більш довгого часу на тренування моделей. В даній роботі для тренування класифікаційних моделей було використано в тому числі і нейронні мережі.

За відсутності достатньої кількості тренувальних даних для будування нейронних мереж, альтернативними методами є логістична регресія та НКБ, що потребують менший обсяг тренувальної вибірки та менше часу для тренування, проте демонструють гірші результати.

У подальших підрозділах буде розглянуто методи класифікації з учителем, що використані у роботі, а також буде розглянуто використання попередньо натренованих моделей. Такі моделі є другою альтернативою для подолання проблеми необхідності промаркованих даних.

Після цього, в підрозділі «Оцінка якості моделей» визначено показники результатів класифікації моделей, за якими буде прийматися рішення щодо доцільності використання попередньо натренованих моделей у порівнянні із традиційними методами машинного навчання.

2.1.1. Логістична регресія

Одним із найпоширеніших методів вирішення задачі класифікації є логістична регресія. Логістична регресія - це статистична модель, що заснована на порівнянні ймовірності настання певної події (залежна змінна), що залежить від множини ознак (незалежні змінні або предиктори), з логістичною кривою. Ця модель застосовується у випадку бінарної залежної змінної, тому її використовують у задачах класифікації. Метод належить до машинного навчання з учителем.

Для опису імовірностей не можна застосовувати звичне рівняння лінійної регресії, оскільки значення можуть виходити за межі множини [0; 1].

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Тому потрібно застосувати обернене логіт-перетворення, яке буде гарантувати, що ймовірність буде у множині [0;1]:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

Метод логістичної регресії став популярним завдяки своїй високій обчислювальній швидкодії та вихідним даним моделі, які допускають оперативну оцінку нових даних, а також можливості роботи як з дискретними, так і неперервними величинами.

Схожим до логістичної регресії за обчислювальною швидкодією є наївний класифікатор Байєса.

2.1.2. Наївний класифікатор Байєса

Наївний класифікатор Байєса – це простий ймовірнісний класифікатор, що заснований на теоремі Байєса з припущенням про незалежність між ознаками. Тобто в основу методу закладено припущення, що поява однієї ознаки ніяк не пов'язана з появою іншої. Саме через це класифікатор називається наївним, оскільки це припущення зазвичай не збігається з істиною. Наївний алгоритм Байєса працює з категоріальними предикторами і результатами.

Формула Байєса, про яку йшлося вище має вигляд:

$$\mathrm{P}(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Спочатку для кожного об'єкта обраховують апостеріорні ймовірності того, що він належить до того чи іншого класу. Апостеріорна ймовірність – умовна ймовірність випадкової події за умови, що відомі апостеріорні дані, тобто отримані після певного досвіду. Об'єкт буде вважатися належним до того класу, до якого апостеріорна ймовірність входження буде максимальною.

Хоча цей метод вважається досить простим, він часто показує достатню точність і часто застосовується у розподілі текстів на два класи. Проте існують такі підходи до обробки природної мови, наприклад, нейронні мережі, що демонструють ще кращу якість моделювання, хоча і мають більш суттєві попередні вимоги.

2.1.3. Використання нейронних мереж для класифікації

Нейронні мережі з прямим зв'язком є універсальним засобом апроксимації функцій. Це дозволяє їх використання для вирішення задач класифікації. Як правило, вони є найефективнішим способом класифікації, бо фактично в нейронних мережах генерується багато регресійних моделей.

Структура нейронної мережі базується на з’єднаних вузлах, які називають нейронами. З’єднанні нейрони здатні передавати сигнал. Нейрон, що отримує

сигнал, здатен обробляти його певним чином, та передавати його до наступних нейронів, що мають із ним зв'язки. Вихід нейрону обчислюється за певною нелінійною функцією суми його входів, наприклад ReLU, sigmoid, та інші.

Зв’язки між нейронами мають певну вагу, що змінюється під час навчання. Завданням процесу тренування є знаходження оптимальних ваг. Ваги впливають на силу сигналу. Для нейрона можна встановити поріг, нижче сигнал не буде сприйматися. Задача бінарної класифікації вирішується мережею з одним нейроном у вихідному шарі, де нейрон може приймати значення від 0 до 1.

Розглянуті методи машинного навчання – нейронні мережі, логістична регресія, наївний класифікатор Байєса – усі мають спільний недолік. Вони потребують промаркованих даних для тренування моделі. Цей недолік нівелюється за допомогою попередньо натренованих моделей, що розглянуті в наступному підрозділі.

2.2. Використання попередньо натренованих моделей

Як раніше було зазначено, за відсутності промаркованих даних, одним із засобів вирішення задач є машинне навчання без вчителя. Проте низка недоліків навчання без вчителя змушують шукати альтернативні способи подолання проблеми відсутності промаркованих даних.

Однією із альтернатив вирішення задач класифікації за відсутності промаркованих даних може бути використання попередньо натренованих моделей, наприклад, моделей zero-shot classification (ZSC) (класифікація без прикладів).

Основною задачею ZSC моделей є класифікація текстових документів для будь-якої множини класів користувача, включаючи класи що не використовувалися під час тренування. Модель може бути створеною один раз, а використовуватися для класифікації до необмеженої множини класів. Звернемо увагу на відмінність від традиційного підходу, де для побудови класифікаційних моделей дослідник має натренувати модель працювати із вичерпним переліком класів, що можуть

використовуватися в майбутньому, а при виявленні додаткового класу, модель повинна бути перенавчена. ZSC моделі допомагають подолати необхідність постійного тренування. І на відміну від стандартного узагальнення в машинному навчанні, де очікується, що класифікатори правильно класифікують нові зразки до класів, які вони вже спостерігали під час навчання, у zero-shot моделях під час навчання класифікатора не надається жодних зразків із класів.

В основі ZSC моделей лежить задача визначення сумісності двох текстів: чи вони протирічать один одному, чи є семантично схожими, чи є нейтральними. Такий підхід був адаптований для задач ZSC: дослідник подає до моделі два текстових документи будь-якої довжини, а модель повертає число від нуля до одиниці, де одиниця означає семантичну суміжність, а нуль – не суміжність.

Перша стаття про zero-shot learning для NLP була представлена у 2008 році на конференції AAAI'08, але назва для нової парадигми навчання була надана «класифікація без даних» [1]. Прорив у роботі із ZSC стався у 2019 році після публікації роботи від Wenpeng Yin. Було запропоновано метод використання попередньо навчених моделей NLI (Natural Language Inference) в якості готових zero-shot класифікаторів. Метод працює шляхом визначення послідовності, яку слід класифікувати як передумову NLI, і побудови гіпотези з кожної мітки-кандидата. Наприклад, якщо для оцінки, чи належить певний текст до класу «*politics*», можна побудувати гіпотезу «*This text is about politics*». Імовірності узгодження і протиріччя гіпотези та тексту потім перетворюються на ймовірності результату класифікації [4].

Використання ZSC моделей супроводжується низкою переваг у порівнянні із традиційними методами машинного навчання із вчителем:

- Уникання процесу маркування даних для етапу тренування;
- Відсутність витрат, пов'язаних із тренуванням моделей, як-то купівля/оренда обчислювальних потужностей, оплата праці інженерів, тощо;
- Нівелюють потенційні помилки серед промаркованих даних можуть зменшують якість традиційних моделей, проте не впливають на якість

класифікації із zero-shot моделями (через відсутність промаркованих даних як таких);

- Нівелюють потенційні помилки, допущені інженерами під час процесу тренування моделей;
- Можливість повторного використання моделі для різних наборів цільових класів без необхідності навчати модель;
- Можливість обробки текстів різними мовами (для окремих моделей).

До недоліків використання ZSC моделей, у порівнянні із традиційним підходом, можна віднести:

- Результати класифікації залежать від назви класу, що подається до моделі. На відміну від традиційного підходу, де назва цільового класу не має значення, і навіть замінюється на числа, при використанні ZSC дослідник має ретельно підбирати слово чи словосполучення для цільових класів.
- В момент обчислення результату класифікації ZSC моделі потребують більше часу, у порівнянні із традиційними методами класифікації. Проте в багатьох випадках така ціна є невеликою у порівнянні із витратами на маркуванні даних та будуванні власних моделей.
- Цільові класи можуть бути зовсім невідомими для моделі, що використовується.

Виходячи з вимог, що висуваються до ZSC моделей, такі моделі потребують надвеликої кількості тренувальних текстових документів, а також тривалого часу для тренування. З огляду на це, ми не будемо на даному етапі наших досліджень будувати власні моделі для ZSC, а будемо використовувати вже натреновані моделі.

В роботі було перевірено, наскільки якісно такі моделі зможуть виконати задачу класифікації текстів; порівняно точність, влучність та чутливість класифікації із використанням навчання зі вчителем (за допомогою нейронних мереж, логістичної регресії, наївного класифікатору Байєса).

В роботі порівняно роботу кількох попередньо натренованих zero-shot моделей, а саме:

- facebook/bart-large-mnli,
- typeform/distilbert-base-uncased-mnli,
- valhalla/distilbart-mnli-12-6,
- valhalla/distilbart-mnli-12-1,
- typeform/roberta-large-mnli.

Модель BART є трансформером, що використовує архітектуру кодувальника-кодувальника із двостороннім кодувальником та авторегресійним декодувальником [12].

Модель distilbert-base-uncased-mnli є версією моделі DistilBERT, що була дотренована спеціально для задач ZSC на базі даних Multi-Genre Natural Language Inference (MNLI), із вимогою нечутливості до регістру. DistilBERT, в свою чергу є дистильованою версією від моделі BERT [13].

Моделі distilbart-mnli-12-1 та distilbart-mnli-12-6 є дистильованими версіями моделі bart-large-mnli, яку було розглянуто вище. Ці моделі створено за допомогою техніки «дистиляції без вчителя». Для моделі distilbart-mnli-12-1 кодувальник мав 12 шарів, а декодувальник – 1 шар; для моделі distilbart-mnli-12-6 кодувальник мав 12 шарів, а декодувальник – 6 шарів.

ZSC модель typeform/roberta-large-mnli побудована на основі загальної NLP моделі RoBERTa [14].

Для оцінки доцільності використання попередньо натренованих моделей необхідно провести класифікацію текстів із промаркованими даними цільових класів. Потім за допомогою показників, що описані в наступному підрозділі, порівняти якість класифікації за допомогою різних підходів: власних моделей та попередньо натренованих моделей.

2.3. Оцінка якості моделей

Після завершення процесу побудови та навчання моделей, необхідно порівняти їх результати. Для порівняння ефективності моделей ми будемо використовувати точність (accuracy), влучність (precision) та чутливість (recall).

Точність – частка результатів, що були правильно класифіковані.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

де TP – (True Positive) це кількість об’єктів позитивного класу, що були коректно класифіковані,

TN – (True Negative) це кількість об’єктів негативного класу, що були коректно класифіковані,

FP – (False Positive) це кількість об’єктів позитивного класу, що були некоректно класифіковані,

FN – (False Negative) це кількість об’єктів негативного класу, що були некоректно класифіковані.

До показнику точності потрібно відноситися із застереженням. У випадку незбалансованих класів (різної кількості спостережень у різних класах) інтерпретація результатів лише за точністю може привести до хибних висновків.

Для коректної інтерпретації результатів розрахуємо влучність та чутливість класифікації. Влучність – частка релевантних спостережень серед тих, що були класифіковані як позитивний клас.

$$Precision = \frac{TP}{TP + FP}$$

Чутливість – частка знайдених спостережень позитивного класу серед усіх наявних спостережень позитивного класу.

$$Recall = \frac{TP}{TP + FN}$$

Використання влучності та чутливості, поряд із точністю, дозволить нам більш коректно інтерпретувати результати класифікації та зробити висновок про доцільність використання того чи іншого методу класифікації.

РОЗДІЛ 3. РОЗРОБКА МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ ТЕКСТІВ ТА ВИКОРИСТАННЯ ПОПЕРЕДНЬО НАТРЕНОВАНИХ МОДЕЛЕЙ

3.1. Збір корпусів текстів для моделювання

Для тренування моделей класифікації у роботі було використано три корпуси текстів із текстами-описами діяльності компаній. Кожен із корпусів був сформований з частини випадкових позицій у базі даних Crunchbase, що знаходяться у відкритому доступі [7]. У цій базі даних є два типи текстів: короткий і довгий описи, де короткий складається з одного речення, а довгий з чотирьох-шести. За умови наявності обох варіантів, їх було об’єднано, в іншому випадку залишався лише короткий опис.

Кожен корпус текстів було розділено на 2 класи. Перший корпус – на класи «є описом компанії, що належить до сектору охорони здоров’я» та «не є описом компанії, що не належить до сектору охорони здоров’я». Другий – «є описом компанії, що виробляє нові технології або надає послуги в стоматологічній галузі» та «не є описом компанії, що виробляє нові технології або надає послуги в стоматологічній галузі». Третій – «є описом компанії, що належить до медичного страхування» та «є описом компанії, що не належить до медичного страхування».

Перший корпус текстів, про належність компаній до сектору охорони здоров’я, складався із 12315 рядків та двох стовпців. Розподіл цільової змінної мав наступний вигляд: 7234 рядків (58,74%) мали значення 1 – компанія належала цільовому сектору; 5081 рядків (41,26%) мали значення 0 – компанія не належала цільовому сектору (рис. 1).

Рис. 1. Розподіл першого корпусу текстів за цільовими класами.

Другий корпус текстів складався із 9427 рядків (рис. 2). Позитивний клас – 2071 рядків (21,97%), негативний клас – 7356 рядків (78,03%).

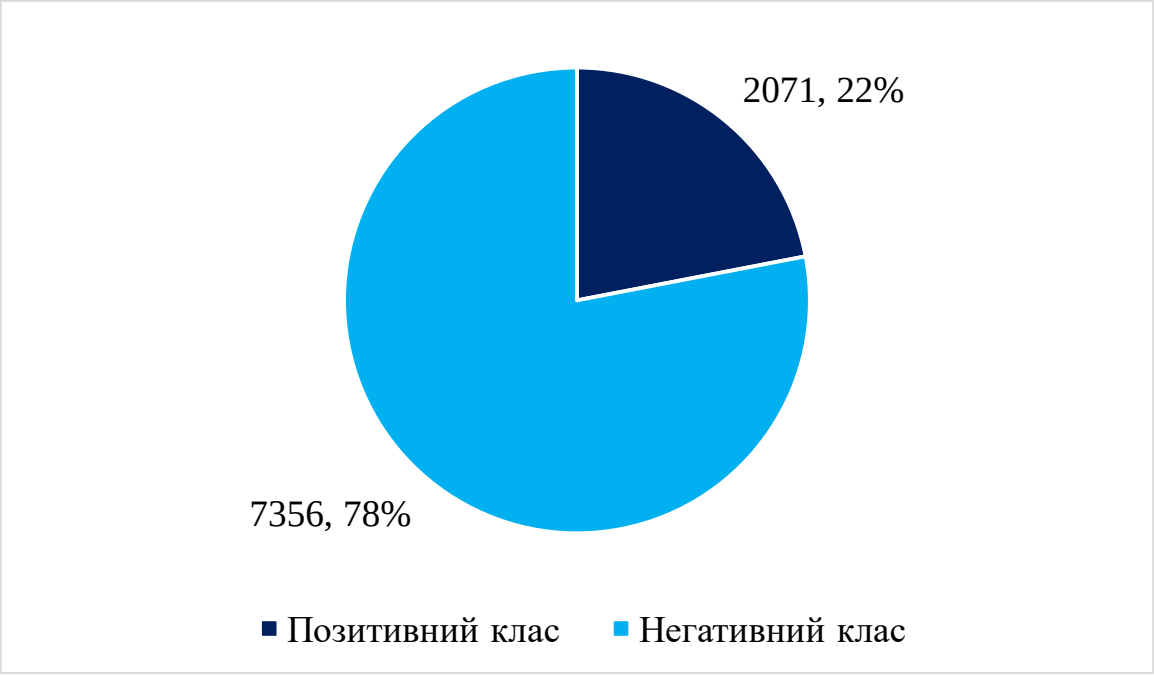

Рис. 2. Розподіл другого корпусу текстів за цільовими класами.

Третій корпус текстів складався із 8761 рядків. Позитивний клас – 3110 рядків (35,50%), негативний клас – 5651 рядків (64,50%).

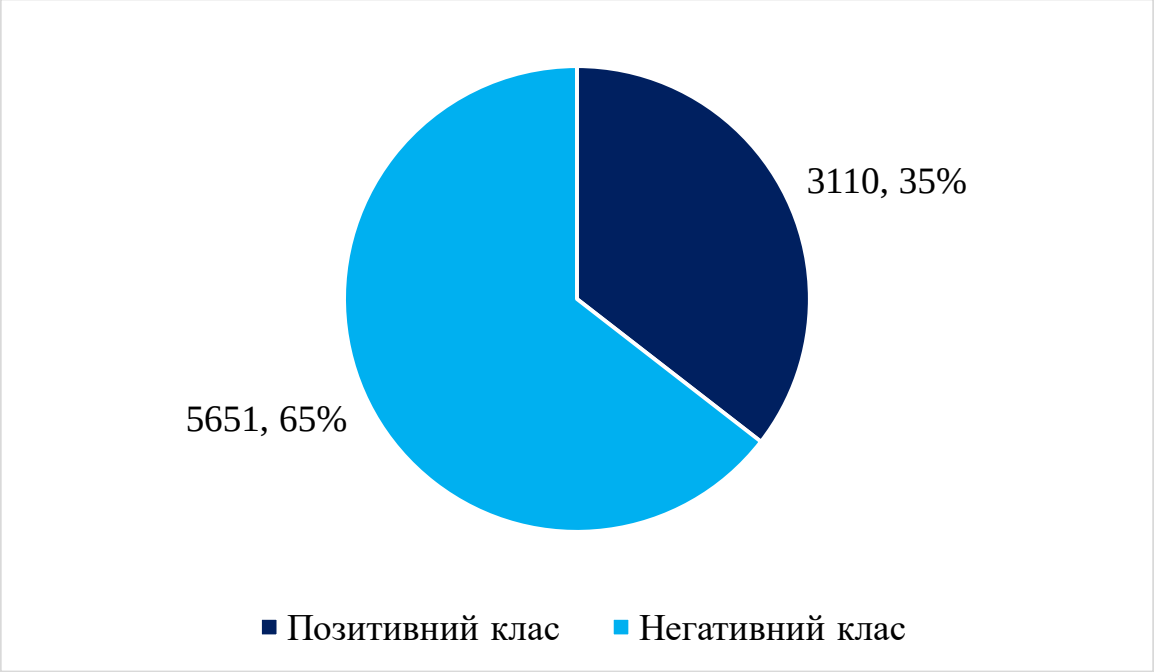

Рис. 3. Розподіл третього корпусу текстів за цільовими класами.

3.2. Первинна обробка тексту

Для проведення класифікації за допомогою методів наївного класифікатора Байєса та логістичної регресії, вільний текст було перетворено до формату bag of words, описаного вище. Для використання цього підходу необхідною умовою є побудова словника слів, що будуть використані під час тренування. Словник слів є сукупністю всіх слів, що будуть використовуватися в процесі тренування моделей. Кількість слів в словниках в усіх випадках дорівнювала загальній кількості унікальних слів, розмір словника не обмежувався. Стоп-слова не використовувалися. Текст був переведений до нижнього регістру. Токенізація тексту та підрахунок кількості слів відбувався за допомогою python бібліотек, а саме sklearn.feature_extraction.text [15].

Для проведення класифікації за допомогою нейронних мереж, перетворення тексту було зроблено за допомогою векторного представлення слів (word embeddings). Текст був переведений до нижнього регістру. Розмір словника складав 2000 найбільш вживаних слів, стоп-слова не використовувалися. Токенізація тексту відбувалася за допомогою python бібліотеки tf.keras.preprocessing.text [16].

3.3. Класифікація за допомогою власних моделей машинного навчання

Для кожного із трьох корпусів текстів побудуємо по три моделі машинного навчання із вчителем: нейронну мережу, логістичну регресію, та наївний класифікатор Байєса (НКБ). Розглянемо результати класифікації для кожного випадку.

Результати класифікації тестової вибірки для першого корпусу текстів представлені в табл. 1.

Таблиця 1

Результати класифікації за допомогою власноруч натренованих моделей для першого корпусу текстів

Модель	Точність	Влучність	Чутливість
Нейронна мережа	0.9728	0.9685	0.9864
Логістична регресія	0.9695	0.9735	0.9755
Наївний класифікатор Байєса	0.9367	0.9254	0.9721

Джерело: розрахунки автора

Найвищу точність та чутливість демонструє нейронна мережа, найвища чутливість у логістичної регресії, а НКБ демонструє найгірші результати серед трьох моделей.

Розглянемо результати класифікації на тестовій вибірці другого корпусу тестів (табл. 2).

Таблиця 2

Результати класифікації за допомогою власноруч натренованих моделей для другого корпусу текстів

Модель	Точність	Влучність	Чутливість
Нейронна мережа	0.9692	0.9462	0.8756
Логістична регресія	0.9691	0.9013	0.7098
Наївний класифікатор Байєса	0.9600	0.8556	0.6295

Джерело: розрахунки автора

Для другого корпусу текстів кращі результати за трьома показниками демонструє нейронна мережа, найгірші – НКБ.

Зрештою, розглянемо результати класифікації на тестовій вибірці третього корпусу тестів (табл. 3).

Таблиця 3

Результати класифікації за допомогою власноруч натренованих моделей для третього корпусу текстів

Модель	Точність	Влучність	Чутливість
Нейронна мережа	0.9070	0.9000	0.8433
Логістична регресія	0.9081	0.9268	0.8172
Наївний класифікатор Байєса	0.8933	0.8431	0.8756

Джерело: розрахунки автора

Для третього корпусу текстів бачимо, що за точністю та влучністю найкращі результати демонструє логістична регресія, а за чутливістю – НКБ.

В усіх трьох корпусах логістична регресія поступається нейронній мережі та чутливістю класифікації, тобто логістична регресія генерує відносно більшу частку хибно негативних класифікацій. НКБ, в свою чергу демонструє найгірші результати за влучністю класифікації, тобто генерує відносно більшу частку хибно позитивних класифікацій. Нейронна мережа загалом демонструє найкращі результати у порівнянні із логістичною регресією та НКБ.

3.4. Класифікація за допомогою попередньо натренованих моделей

Оцінимо можливості класифікації корпусів текстів за допомогою попередньо натренованих zero-shot classification (ZSC) моделей. Цей підхід зробить можливим без використання тренувальних вибірок зробити оцінку, чи відповідає певний конкретний текстовий документ, наприклад, опису компанії, що належить до сектору охорони здоров'я. Для п'яти моделей було оцінено точність, влучність та чутливість класифікації.

Результати класифікації для першого корпусу текстів представлені в табл. 4.

Таблиця 4

Результати класифікації за допомогою наперед натренованих моделей для першого корпусу текстів

Модель	Точність	Влучність	Чутливість
facebook/bart-large-mnli	0.9643	0.9982	0.9408
typeform/distilbert-base-uncased-mnli	0.9297	0.9194	0.9649
valhalla/distilbart-mnli-12-6	0.9275	0.9084	0.9748
valhalla/distilbart-mnli-12-1	0.9140	0.9385	0.9135
typeform/roberta-large-mnli	0.8775	0.9603	0.8257

Джерело: розрахунки автора

Найкращу точність та влучність класифікації демонструє модель facebook/bart-large-mnli. Найвища чутливість була досягнута моделлю valhalla/distilbart-mnli-12-6, проте вона суттєво поступається за влучністю (90,8% проти 99,8%).

Результати класифікації для другого корпусу текстів представлені в табл. 5. Результати є аналогічними до результатів на першому корпусі. Модель facebook/bart-large-mnli знову демонструє найвищі показники точності та влучності. Модель valhalla/distilbart-mnli-12-6 знову має найвищу влучність, і знову поступається за влучністю (86% проти 91%). Три інші моделі демонструють гірші результати.

Таблиця 5

Результати класифікації за допомогою попередньо натренованих моделей для другого корпусу текстів

Модель	Точність	Влучність	Чутливість
facebook/bart-large-mnli	0.9667	0.9103	0.9411
typeform/distilbert-base-uncased-mnli	0.8373	0.6346	0.6122
valhalla/distilbart-mnli-12-6	0.9565	0.8616	0.9556
valhalla/distilbart-mnli-12-1	0.8990	0.7162	0.8947
typeform/roberta-large-mnli	0.9524	0.9076	0.8725

Джерело: розрахунки автора

Розглянемо результати класифікації для третього корпусу текстів (табл. 6).

Таблиця 6

Результати класифікації за допомогою наперед натренованих моделей для третього корпусу текстів

Модель	Точність	Влучність	Чутливість
facebook/bart-large-mnli	0.9035	0.9354	0.7823
typeform/distilbert-base-uncased-mnli	0.7963	0.7192	0.6990
valhalla/distilbart-mnli-12-6	0.8961	0.8501	0.8588
valhalla/distilbart-mnli-12-1	0.8185	0.7394	0.7547
typeform/roberta-large-mnli	0.8873	0.9303	0.7379

Джерело: розрахунки автора

Результати класифікації є аналогічними до результатів на попередніх: facebook/bart-large-mnli та valhalla/distilbart-mnli-12-6 мають найвищу якість у порівнянні із іншими трьома моделями. Модель facebook/bart-large-mnli знову показує найвищий показник влучності, а valhalla/distilbart-mnli-12-6 – найвищій показник чутливості.

3.5. Порівняння якості класифікації за допомогою власних та попередньо натренованих моделей

Порівняємо результати класифікації за допомогою власних моделей із результатами класифікації за допомогою ZSC моделей. Виходячи з того, що серед ZSC моделей найкращі результати на всіх корпусах текстів демонстрували facebook/bart-large-mnli та valhalla/distilbart-mnli-12-6, в цьому підрозділі не ми будемо розглядати інші ZSC моделі.

Аналізуючи результати табл. 1 та табл. 4 бачимо, що на першому корпусі текстів використання моделі facebook/bart-large-mnli призводить лише до незначного падіння точності моделі, у порівнянні із нейронною мережею та логістичною регресією, а у порівнянні із НКБ має навіть вищу точність. На другому корпусі текстів (табл. 2 та табл. 5) відзначаємо відносно вищі показники чутливості для ZSC моделей у порівнянні із власними моделями, а facebook/bart-large-mnli і за влучністю, і за чутливістю краще за логістичну регресію та НКБ. Для третього корпусу текстів (табл.3 та табл. 6) всі п’ять моделей (три власні та дві ZSC мають схожу точність 89-90%, facebook/bart-large-mnli переважає за влучністю, valhalla/distilbart-mnli-12-6 – за точністю.

Проаналізувавши результати трьох експериментів, можемо зробити висновок, що якщо мається достатній обсяг промаркованих даних для побудови нейронної мережі, то нейронна мережа буде показувати кращі результати у порівнянні із попередньо натренованими моделями. Але у випадку недостатньої кількості промаркованих даних, використання наперед натренованих моделей та поєднання їх результатів може бути альтернативним варіантом для досягнення необхідної точності класифікації. Більш того, вимоги до класифікації з боку бізнесу можуть схиляти до вибору тієї чи іншої моделі. Зокрема, у випадку необхідності максимізації влучності класифікації, тобто мінімізації хибно позитивних випадків, доцільним є використання моделі facebook/bart-large-mnli. У випадку необхідності максимізації чутливості, тобто мінімізації хибно негативних випадків, доцільним є використання моделі valhalla/distilbart-mnli-12-6.

ВИСНОВКИ

В рамках роботи було зібрано корпуси текстів-описів компаній біотехнологічного напрямку та підготовлено їх до застосування моделей машинного навчання за допомогою методів обробки природної мови.

Були розглянуті традиційні методи класифікації тексту за допомогою машинного навчання, та були побудовані власні класифікатори на основі нейронної мережі, логістичної регресії та наївного класифікатора Байєса. Для кожного із цих методів, було побудовано моделі для різних корпусів текстів. Відзначаємо, що серед цих трьох методів найкращі результати демонстрували моделі, побудовані за допомогою нейронних мереж.

В якості альтернативи до моделей із вчителем та моделей без вчителя, було розглянуто спосіб класифікації текстів із використанням попередньо натренованих zero-shot classification (ZSC) моделей, що дозволяє

- уникнути етапу маркування даних;
- уникнути етапу тренування моделі;
- проводити класифікацію текстів навіть за тими класами, що не використовувалися під час тренування моделі.

Було зроблено класифікацію трьох корпусів текстів за допомогою п’яти різних попередньо натренованих ZSC моделей, та було порівняно якість класифікації за допомогою цих моделей із якістю класифікації за допомогою традиційних методів машинного навчання, як-то нейронна мережа, логістична регресія та наївний класифікатор Байєса.

За результатами такого порівняння було зроблено висновок, що деякі ZSC моделі, наприклад, facebook/bart-large-mnli та valhalla/distilbart-mnli-12-6, лише несуттєво поступаються за якістю традиційним методам, натренованих на власних тренувальних корпусах текстів, навіть коли наявні тексти містять професійні слова та терміни.

Більш того, у випадку різної ціни помилки хибно-позитивних та хибно-негативних випадків, користувачу може бути доцільним розглядати різні

попередньо натреновані моделі. Так facebook/bart-large-mnli в трьох випадках мала найкращі показники влучності, а valhalla/distilbart-mnli-12-6 – найкращі показники чутливості класифікації.

В той же час, використання нейронних мереж приведе до більш якісної класифікації, у порівнянні як з іншими традиційними методами, так і у порівнянні з попередньо натренованими моделями, за умови наявності достатнього розміру корпусу текстів для побудови нейронної мережі.

Перспективними напрямами подальших досліджень є аналіз доцільності донавчання існуючих попередньо натренованих моделей для введення до тренувальної множини спеціальних професійних термінів біотехнологічного напрямку, та створення нових zero-shot classification моделей, натренованих із самого початку виключно на текстах із спеціальними термінами.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Chang et al. Importance of semantic representation: Dataless classification. 2008. In AAAI, 830–835. [Електронний ресурс]. – Режим доступу: https://www.aaai.org/Papers/AAAI/2008/AAAI08-132.pdf
2. Srivastava et al. Zero-shot learning of classifiers from natural language quantification. 2018. [Електронний ресурс] – Режим доступу: https://aclanthology.org/P18-1029
3. Obamuyide A., Vlachos A. Zeroshot relation classification as textual entailment. 2018. Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 72–78. [Електронний ресурс] – Режим доступу: https://aclanthology.org/W18-5511
4. Yin et al. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach, 2019. [Електронний ресурс] – Режим доступу: https://arxiv.org/abs/1909.00161
5. Офіційний сайт Hugging Face. [Електронний ресурс]. – Режим доступу: https://huggingface.co/
6. Wolf et al. Transformers: State-of-the-Art Natural Language Processing // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. – 2020 – С. 38-45. [Електронний ресурс]. – Режим доступу: https://www.aclweb.org/anthology/2020.emnlp-demos.6
7. Офіційний сайт Crunchbase. [Електронний ресурс]. – Режим доступу: https://www.crunchbase.com
8. Rebecca Lake. A Guide to the 11 Market Sectors. [Електронний ресурс] – Режим доступу:https://finance.yahoo.com/news/guide-11-market-sectors-142851510.html
9. Smiljanic Stasha. The State of Healthcare Industry – Statistics for 2022. [Електронний ресурс]. – Режим доступу: https://policyadvice.net/insurance/insights/healthcare-statistics

10. Офіційний сайт Всесвітньої організації охорони здоров'я. [Електронний ресурс]. – Режим доступу: https://www.who.int
11. Bag of Tricks for Efficient Text Classification [Електронний ресурс] / A.Joulin, E. Grave, P. Bojanowski, T. Mikolov – Режим доступу: https://arxiv.org/pdf/1607.01759.pdfi
12. Lewis et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension [Електронний ресурс]. – Режим доступу:https://arxiv.org/abs/1910.13461
13. Devlin et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/1810.04805v2
14. Liu et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/1907.11692
15. Документація python-бібліотеки scikit-learn [Електронний ресурс] – Режим доступу: https://scikit-learn.org/
16. Документація python-бібліотеки TensorFlow [Електронний ресурс] – Режим доступу: https://www.tensorflow.org/