Міністерство освіти і науки України

Національний університет «Києво - Могилянська академія»

Факультет соціальних наук та соціальних технологій

Кафедра соціології

Кваліфікаційна робота

освітній ступінь–бакалавр

на тему:

«ІНТЕГРУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ У СУЧАСНУ СОЦІОЛОГІЧНУ МЕТОДОЛОГІЮ В УКРАЇНІ»

Виконала: студентка 4 - го

року навчання спеціальності

054 «Соціологія»

Прокудіна Маргарита Олександрівна

Керівник: Артикуца С.С.
старший викладач кафедри соціології НаУКМА

Рецензент:

Кваліфікаційна робота захищена

з оцінкою «_____»

Секретар ЕК: _____

«___» _____ 2020 р.

Київ–2020

ЗМІСТ

ВСТУП

Протягом останніх двох десятиліть генерація великих масивів даних почалася з новою силою, адже у соціальний простір втрутилися багатофункціональні гаджети з підключенням до Інтернету.

Згідно до прогнозів дослідників компанії IDS, які були зроблені у доповіді «IDC`s Data Age 2025 study, sponsored by Seagate» (IDC, 2017) до 2025 року середньостатистична людина буде взаємодіяти зі своїм гаджето м до 4800 разів на день, а це приблизно кожні 18 секунд. Також за даними ID S ми можемо прослідкувати динаміку накопичення необроблених статистичн их даних. Станом на 2003 рік у світі вже було 5 ексабайтів даних (1 ЕБ = 1 мл рд гігабайтів). У 2008 році людство нараховувало близько 0,18 зеттабайтів (1 ЗБ = 1024 ексабайтів), а у 2015 році було накоплено уже 6,5 зеттабайтів дани х. До того ж, до 2020 року людство зможе розпоряджатися
вже 40 - 44 зеттабайтами інформації, А вже до 2025 цей показник виросте у 1 0 разів.

Подібна висока залученість сервісів збирання інформації до рутинних практик індивідів спровокувало появу нового типу даних – “big data” або великих даних. У свою чергу, це підвищило вірогідність використання методів обробки цих даних у сучасній соціологічній методології, адже соціологія як дисципліна, що вивчає взаємодію у суспільстві та масові соціальні практики, напряму зацікавлена в обробці та інтерпретації подібних даних.

Проте перед дослідником постають чотири головні проблеми обробки цих даних. Першою є невизначеність для яких завдань можна використовувати методи машинного навчання у соціології. Друга - складність вибору самого методу для поставленої задачі. Третя - складність пошуку чистого масиву інформації. І остання - нерозуміння

як практично втілити алгоритм у життя.

Мета дослідження: сформувати модель інтегрування методів машинного навчання в соціологічну методологію в Україні.

Об'єктом дослідження виступили безпосередньо методи машинного навчання, та *предметом* – їх інтегрування у сучасну соціологічну методологі ю.

Задля досягнення кінцевої мети було встановлено такі *завдання:*

1. Систематизувати категоріально - понятійний апарат методів машинного навчання;
2. З'ясувати доцільність інтегрування методів з інших дисциплін у соціологію;
3. Виявити перелік методів машинного навчання, які доцільно інтегрувати у соціологію;
4. Визначити основні проблеми застосування методів машинного навчання у соціологічну методологію;
5. Окреслити основні моделі застосування машинного навчання в соціологічній методології в Україні.

Для більшої достовірності теоретичних висновків та припущень а також для забезпечення коректності екстраполювання їх з західної практики на українську, у роботі використовується метод експертних інтерв'ю з провідними спеціалістами кожної з зазначених тематик. Експерти поділені на дві категорії: теоретичного підґрунтя та практичного застосування. Перша категорія експертів являє собою фахівців з соціологічної методології, які були опитані першими та на основі відповідей яких формується перелік методів машинного навчання, які мають найбільші

перспективи інтеграції в соціологічну методологію. На основі цього переліку був підібраний список експертів з другої категорії - фахівців с певного методу машинного навчання, що так чи інакше пов'язаний з вирішенням соціологічних завдань. Такий метод відбору респондентів дозволяє всебічно дослідити об'єкт та предмет дослідження та уникнути випадкових помилок у інтерпретуванні результатів інтерв'ювання.

Усі інтерв'ю є напів структурованими, адже торкаються не тільки теоретичних та практичних засад роботи машинного навчання, а й особистого досвіду його використання експертами. Також зважаючи на специфіку кожної категорії експертів, для кожної з них був розроблений окремий гайд (Додаток А, Додаток Б). Емпірична частина полягає в аналізі 5 інтерв'ю. Принцип відбору респондентів для першої категорії полягав у обгрунтуванні експертності за стажем роботи з подібною тематикою, кількість фахових наукових публікацій а також загальний науковий інтерес у методах машинного навчання. Принцип відбору респондентів для другої категорії засновувався на аргументуванні їх практичного досвіду роботи з певними методами машинного навчання та інтеграції цих методів для вирішення певних соціологічних задач.

РОЗДІЛ 1. ПРАКТИКА ВИКОРИСТАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ В УКРАЇНІ ТА ПЕРСПЕКТИВИ ЇХ ІНТЕГРУВАННЯ В СОЦІОЛОГІЮ

1.1. Поняття «data science» та категоріально - понятійний апарат машинного навчання

Математичні здібності людського мозку на даному етапі еволюції є обмеженими, тому він не здатен обробити величезні масиви інформації за короткий проміжок часу. У даному випадку його може обійти навіть звичайний калькулятор. Якщо провести експеримент та попросити будь-яку пересічну

людину помножити число 258 на 179, є велика вірогідність того, що вона навіть не намагатиметься це зробити (хоча це і досить осяжне завдання). Але незважаючи на це, людський мозок має одну перевагу над вимірювальними машинами – він підлаштовується під зміни соціального середовища. «Data Science» займається тим, що мінімізує дані розбіжності шляхом впровадження прогресивних методів обробки великих масивів інформації – шляхом машинного навчання. І головною перевагою інтеграції методів машинного навчання у сучасну соціологічну методологію є те, що соціологи зможуть користуватися не тільки раніше недоступними джерелами (і об’ємами цих джерел) даних, а й новим методологічним підходом впорядкування характеристик явищ, що досліджуються.

Не дивлячись на те що «на слуху» наука про великі дані тільки останні півтора десятиліття, перші згадки про неї простежуються ще у 50 - их роках. І не дивлячись на досить невеликий час перебування у полі зору наукової спільноти, за даними порталу з моніторингу ситуації на ринку праці США Glassdoor, станом на 2019 рік (25 Highest Paying Jobs in America,

2019) професії, пов’язані з аналітикою великих даних, входить в топ - 25 най більш високооплачуваних професій.

До того ж, якщо орієнтуватися на результати аналізу сервісу Google Trends, то кількість запитів, а отже і інтерес до такого явища як data science, почав зростати починаючи з 2014 року:

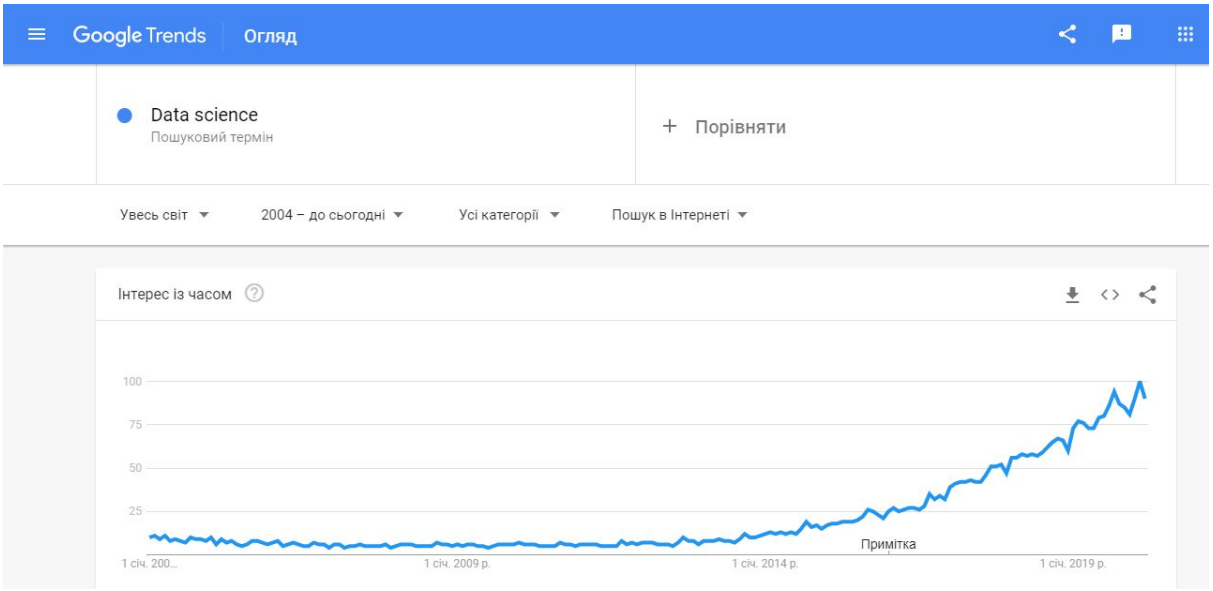

Рис. 1.1. **Дані за 2004 - 2020 роки з сервісу Google Trends по запиту «Data science»**

Загалом наука «data science», незважаючи на свій молодий вік, є розвиненою та розгалуженою. Лише за свою коротку історію вона мала декілька хвиль теоретичних та практичних злетів та падінь (Tilburg University, 2019). Технологічне становлення сучасної парадигми не відбувалось лінійно, та розвиток її припадав саме на моменти дефіциту розвинутого інструментарію. Так, після усвідомлення потреби автоматизування та масштабування завдань класифікації та регресії, виникли перші, найпростіші алгоритми машинного навчання. Але з часом та розвитком теоретичного та практичного знання в

області data science, ці методи доповнювались, з’являлись кардинально нові, до них створювався новий інструментарій і наразі ми маємо досить складну т а розгалужену схему, яка потребує чіткої класифікації. Саме тому є доречни м окреслити базову класифікацію методів машинного навчання та дати харак теристику кожному з них. Для простішого розуміння наводимо зручну схему методів машинного навчання, включно з окремими алгоритмами.

Таблиця 1.1

Схема алгоритмів машинного навчання

Машинне навчання				
Класичні методи		**З підкріпленням**	**Ансамблеві методи**	**Нейромережі**
Без вчителя	**З вчителем**			
Кластеризація	Класифікація		Стекінг	CNN
Пошук правил	Регресія		Беггінг	RNN
Зменшення розмірності			Бустінг	GAN
				Автокодувальник
				Перцептрони

Головне завдання машинного навчання – це передбачити результат на осн ові певного набору даних, які надав дослідник. (Cielen D., Arno D. B., Meysman, Ali M., 2016). Головними його складовими є:

Масив даних – чим більший і різноманітніший буде масив даних, тим біль ш точним буде результат. На основі цього масиву даних (важливо, аби він бу в «чистий») у подальшому алгоритм буде навчатися та будувати закономірно сті, виконувати свою предикативну функцію, тощо. (EMC Education

Services, 2015)

Зазвичай коли говорять про великі дані у контексті соціологічних дослідж ень, мається на увазі десятки тисяч респондентів. У випадку з «big data» необ хідно розуміти, що ці дані мають налічувати мільйони, мільярди, міріади відп овідей. Тільки така кількість забезпечить максимально точні результати та до зволить використовувати функціонал великих даних у повному обсязі.

Дані бувають структурованими та неструктурованими. Структуровані – ц е дані, що зберігаються у замовника у вже визначеному форматі (таблиці, схе ми тощо), а неструктуровані це, відповідно, документи що не мають чіткого в нутрішнього структурування та класифікації.

Звідки взяти необхідні дані у такому обсязі – одна з найбільших проблем для дослідників, адже звичайні методи збору кількісних даних для цього не п ідходять. Протее це питання ми детально розглянемо у наступних розділах.

Другою необхідною складовою машинного навчання є ознаки. Ознаки – ц е класифікація самим дослідником досліджуваного об’єкта за певними характ еристиками. Дані характеристики необхідні для того, аби алгоритм мав вказів ки до класифікування вхідних значень. Але варто зазначити, що має місце і с итуація, коли дослідник у разі ручного вводу класифікаторів привносить у до слідження певну суб’єктивність і з’являється ризик отримання недостовірних результатів.

І третьою необхідною складовою машинного навчання є самі алгоритми. Алгоритмів у темі штучного інтелекту є безліч, і вони відіграють роль певних методів обробки даних (або набору методів). Вони відрізняються між собою завданнями, які запрограмовані вирішувати та точністю і ефективністю вико нання цих завдань.

Перш за все, почати необхідно з класичних методів **машинного навчанн**

я, які поділяються на методи з навчанням за допомогою вчителя (supervised le arning) та без нього (unsupervised learning). (James G., Witten D., Hastie T., Tibshirani R., 2017)

Навчання з вчителем – теоретичні засади цих алгоритмів були створені щ е у середині минулого століття, та представляли собою прості алгоритми пош уку закономірностей.

До цього методу належать алгоритми: K - NN, Наївний Байес, SVM, Дере ва Рішень, Логістична Регресія, Лінійна Регресія, Поліноміальна Регресія.

Наявність вчителя гарантує для алгоритму демонстрацію взірцевих модел ей, на яких вона
може навчатися. Наприклад, учитель показує алгоритму різні варіанти націон ального одягу, алгоритм навчається, і в подальшому може розрізняти вбрання на, припустимо, старих фотографіях.

Цей метод машинного навчання зазвичай використовується у завданнях к ласифікації та регресії. (Annalyn Ng, Kenneth Soo, 2017)

Навчання без вчителя з'явилося набагато пізніше за навчання з вчителем (у 90 - ті роки). На відміну від навчання з вчителем, тут дослідник надає алго ритму велику кількість необроблених даних (у даному випадку – фото людей у різних національних костюмах) та дає йому команду самому встановити кр итерії, за якими він буде їх розмежовувати та класифікувати.

До цього методу належать алгоритми: Agglomerative, DBSCAN, метод k - середніх, MeanShift, Fuzzy C - means, Euclat, Apriori, FP - Growth, y - SNE, P SA, LSA, SVD, LDA.

Цей метод використовують для завдань кластеризації, пошуку закономірн остей та правил, зменшення розмірності.

Крім класичних методів, існують також ще 3 самостійних методи – навча

ння з підкріпленням, ансамблеві методи та нейронні мережі. (Hastie T., Tibshirani R., Friedman J.).

Навчання з підкріпленням є одним з найдалекіших від соціології методі в, але про нього необхідно сказати також. Цей метод засновується не на пред икативній функції алгоритму, а на реагуванні на зовнішні фактори, пошуку н айвдалішого й найефективнішого рішення з усіх. Наприклад, подібний алгор итм зміг у 2016 році обіграти в го одного з найкращих гравців у світі (Artifici al intelligence: Google's AlphaGo beats Go master Lee Se-dol, 2016). Специфіка цієї гри в тому, що продумати всі варіанти ходів наперед не можливо – їх більше, ніж будь - яке з осяжних чисел. І саме ситуативна мане ра поводження є характерною рисою навчання з підкріпленням. У сучасному світі це закладено до принципів роботи роботів - пилососів та у розробці авт омобільних автопілотів.

До цього методу належать алгоритми: генетичний алгоритм, АЗС, SARSA , Q - Learning, Deep Q - Network.

Ансамблеві методи є одними з найпопулярніших та найбільш надійних ме тодів у цілому. Суть його дуже проста – декілька методів виправляють помил ки один одного і на виході дають певне середнє значення, яке буде найточніш им з можливих у моделі. Таким чином, якість моделі, що скомбінована з різн их можливих алгоритмів набагато більша ніж якість просто алгоритмів (а кож ен зі своїми перевагами та недоліками), взятих поодинці. Цікаво, що чим біль ш хиткий алгоритм при внесенні вхідних даних (як звичайна Регресія чи Дере во Рішень), тим краще для саме цієї моделі.

Ансамблеві методи поділяються на три способи реалізації:

Стекінг – найменш популярний з усіх. Головна особливість у тому, що рі зні алгоритми оцінюють вхідні дані за схемою, на яку запрограмовані, а поті

м передають

їх на вхід до останнього алгоритму, який обробляє дані, виводить середнє та безпосередньо приймає рішення.

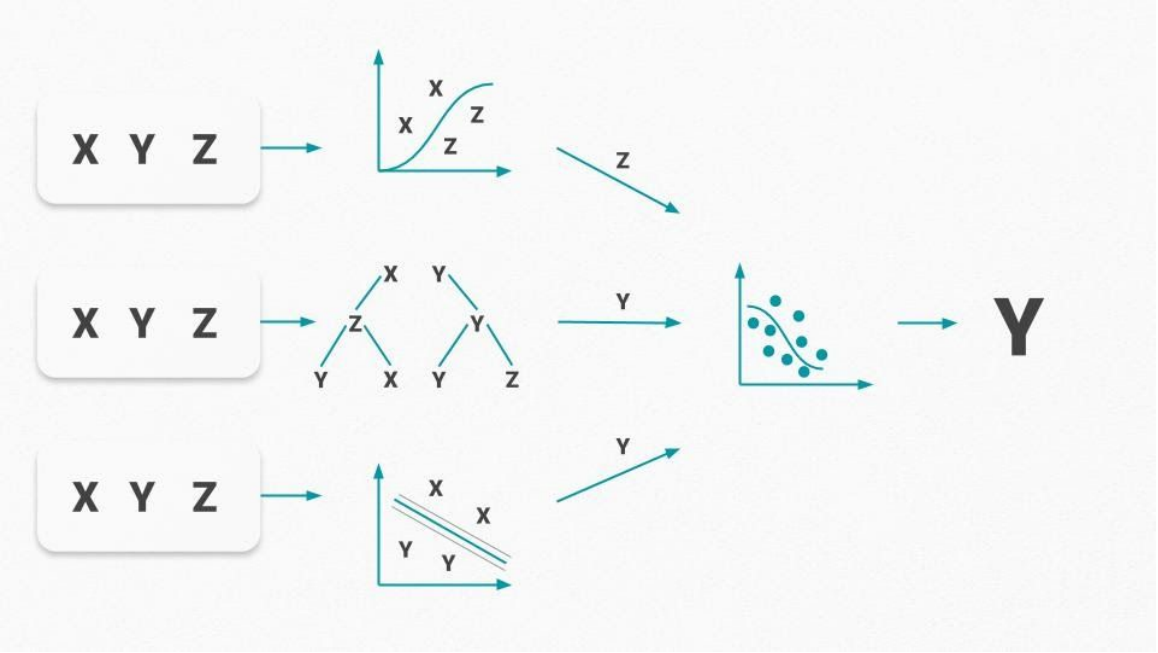

Рис. 1.2. **Схема роботи стекінгу**

Другий спосіб з ансамблевого методу – це *бегінг*, який реалізується завдяк и алгоритму Random Forest. На відміну від стекінгу, він використовує один ал горитм, але проганяє його через декілька різних випадкових вибірок з масиву , а потім методом простого голосування вибирає відповідь-переможця.

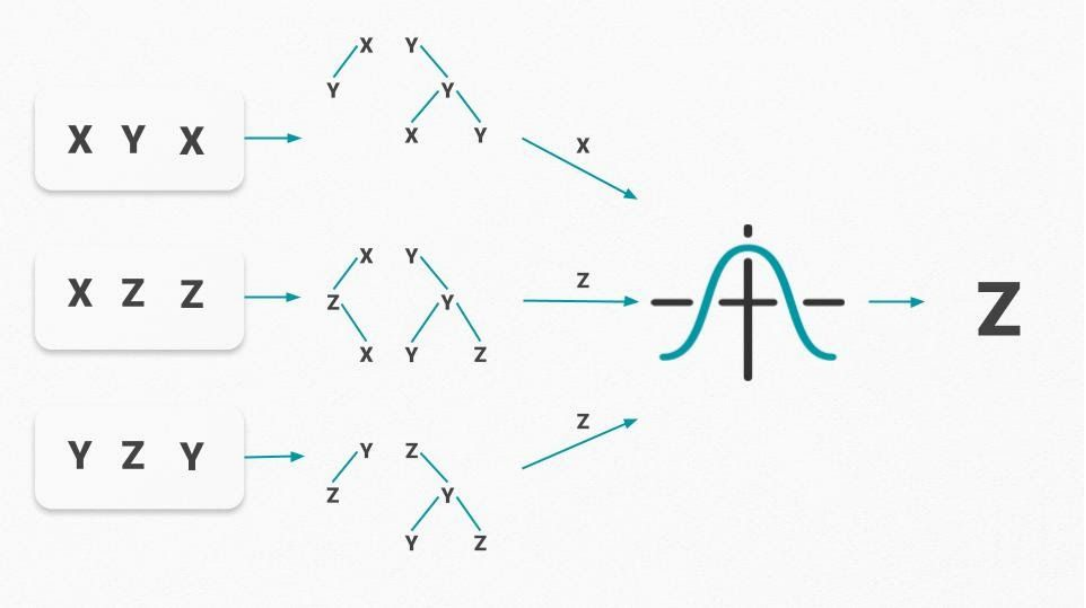

Рис. 1.3. **Схема роботи бегінгу**

Одна з головних переваг бегінгу – це те, що обробка даних йде паралельно, а отже зменшує кількість часу на обробку даних. Саме через цю особливість його використовують частіше, ніж наступний спосіб, адже незважаючи на те, що він не такий точний, здатність паралельного завантаження робить його фаворитом серед технічних систем.

І останній спосіб в ансамблевому методі – це *бустінг*. (Shai Shalev-Shwartz, Shai Ben-David, 2014) Ключова особливість цього алгоритму у тому, що всі його внутрішні алгоритми йдуть послідовно, приділяючи увагу більше не тим даним, які пройшли без ускладнень а тим, на яких були виявлені проблеми. Фактично, дослідник намагається довчити алгоритм на помилках попередніх алгоритмів. Це забезпечує найточніший результат з можливих у цьому методі.

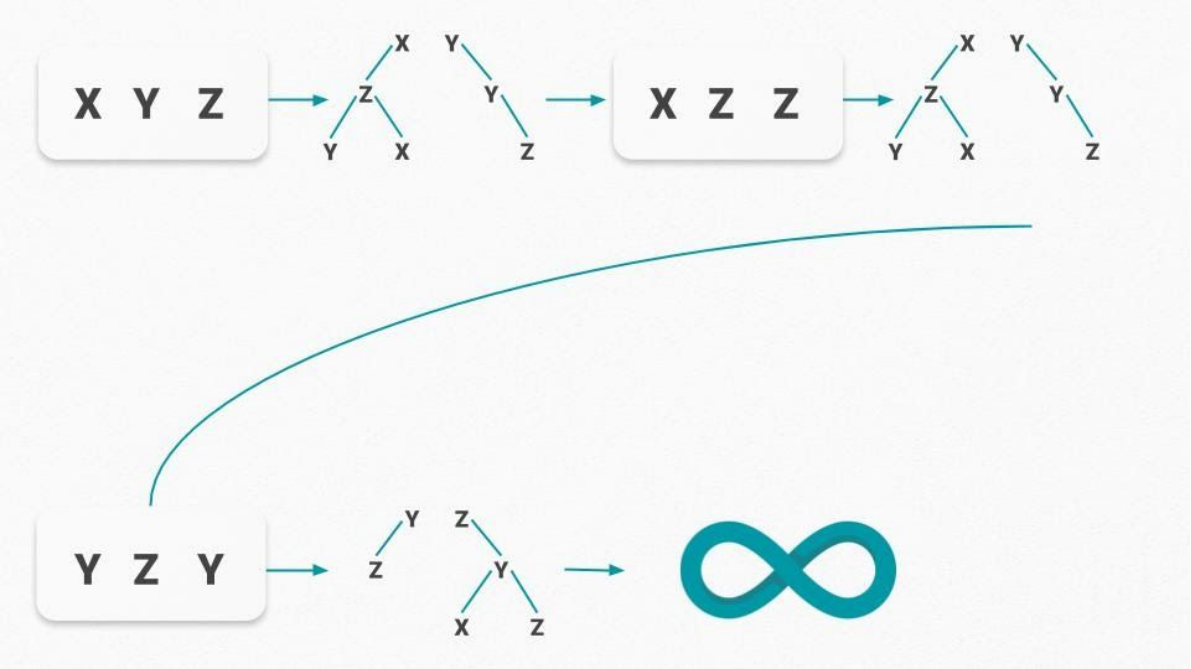

Рис. 1.4. **Схема роботи бустінгу**

До цього способу належать алгоритми: AdaBoost, CatBoost, XGBoost, Ligh tGBM.

Ну і останній з методів машинного навчання – це **метод штучних нейрон них мереж.**

Штучні нейронні мережі відкривають перед сучасними дослідниками можлив ості не тільки до масштабування самих вибірок, але і до змінення вектору дос ліджень у зв’язку з новими типами даних, які можна досліджувати.

Наразі цей метод застосовується замість усіх вищезгаданих методів та у з адачах обробки зображень, розпізнавання та синтезування голосу, виявлення певних об’єктів на фото/відео.

Алгоритм нейронних мереж – це метод, який імітує поведінку нейронних мереж у мозку людини. Він використовується для того, щоб досягти певного результату без залучення великої кількості людських ресурсів а також для авт оматизування певної діяльності.

Нейронні мережі в людському організмі відіграють провідну роль, і саме завдяки їх структурі і були втілені у штучному вигляді. Поляков у своїй праці «О принципах нейронной организации мозга» (
1965) демонструє, що вони слугують агентами певних обчислювальних функ цій у контексті реагування на певні ситуації. Саме вони приймають найбільш вигідні та безпечні для всього організму рішення в рамках певних заданих ум ов. Ці мережі мають входи (дендрити), виходи (аксони), місця з'єднань (сина пси), нейрон та його поріг збудження, який є головним центром прийняття рі шень щодо усіх ситуацій.

Схематично цей процес виглядає так:

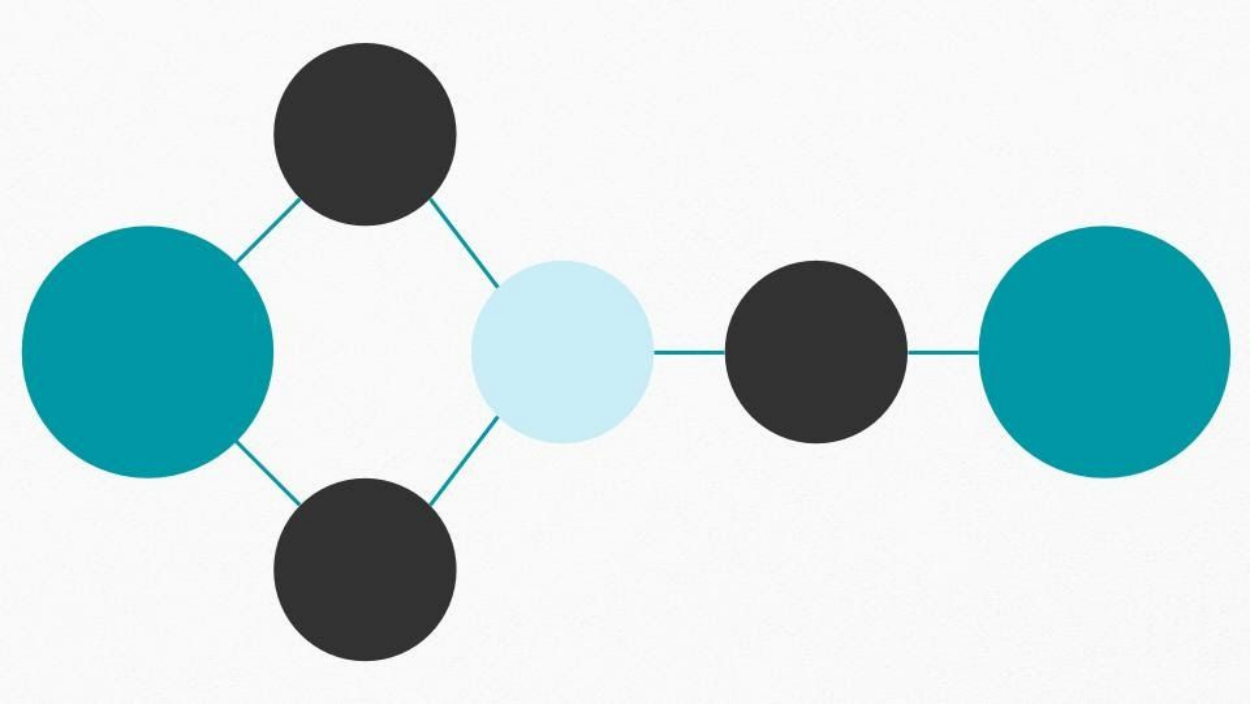

Рис. 1.5. **Схема роботи нейрону у живому організмі**

Коли на вхідний шар нейрону надходить певний імпульс і він є достатнім для проходження порогу збудження нейрону, він передається через аксони до дендритів наступного за ним нейрону.

Якщо ж говорити про штучно створену модель, то метафорично, вона діє

за тією ж логікою (Горбань, 1990):

Вхідний шар – на нього надходить певна кількість неструктурованих дани х;

Прихований шар – це математичний алгоритм, завдяки якому відбуваєтьс я транспортування оброблених значень на наступний шар нейрону. У його ск лад входить дві частини: суматор, який обробляє, класифікує та перераховує значення, які до нього надійшли та нелінійний перетворювач, що являє собо ю функцію активіції нейрону. Саме завдяки ньому приймається рішення, яког о роду дані надходять до вторинного процесу обробки:

Вихідний шар – передає отриману оброблену інформацію як остаточну ві дповідь до зовнішнього середовища.

Схематично, цей процес буде дуже схожим на вищезгадану біологічну мо дель нейронів:

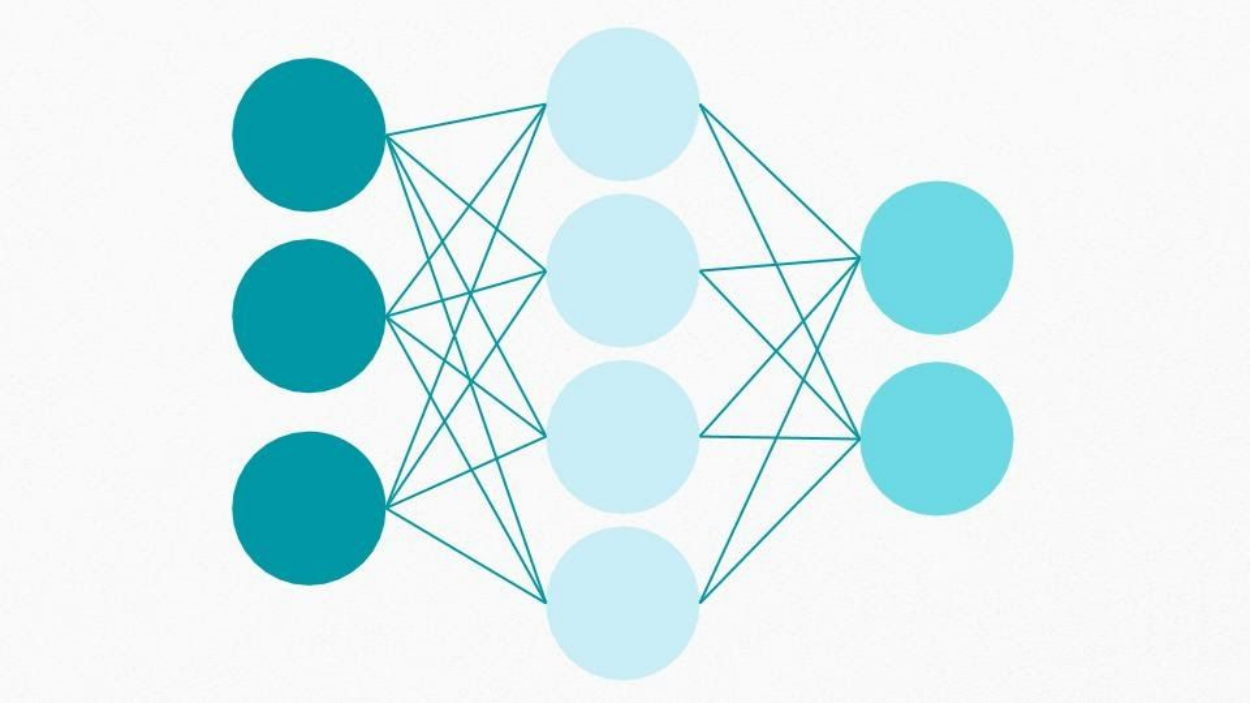

Рис. 1.6. **Схема роботи елементарної штучної нейромережі**

Найпопулярнішою та найпоширенішою є модель (під цим розуміється ана логова модель, яка попереджає бінарні вихідні значення), суматор у вигляді с игмоїдальної функції, що використовує зважену суму всіх коефіцієнтів. Прихованих шарів може бути б езліч, а це означає, що певна кількість вхідних значень може бути перерахова на ними велику кількість разів.

Отже, у роботі було систематизовано категоріально - понятійний апарат машинного
навчання, що виступає досить розгалуженою та багаторівневою наукою. З ог ляду на всю інформацію ми можемо попередньо дійти висновку, що вона мо же бути перспективним напрямком для використання дослідниками - соціол огами та має певні методологічні підвалини до подальшої інтеграції в соціоло гічні методи.

1.2. Методологія дослідження інтегрування методів машинного навчання у сучасну соціологічну методологію в Україні

У даному досліджені використаний метод експертних інтерв'ю. Даний метод аргументований вузькою спеціалізацією у цій темі, та її міждисциплінарним спрямуванням, а отже дана тема не є буденою та знайомою для більшості соціологів, а притаманна більше спеціалістам, що працюють на межі двох дисциплін та мають набір специфічних навичок. Емпірична частина полягає у аналізі п'яти напівструктурованих інтерв'ю. Експерти були поділені на дві категорії, і кожна категорія була відібрана за допомогою певного набору критеріїв та була опитана за своїм гайдом (Додаток А, Додаток Б).

Перша категорія респондентів - це люди з профільною соціологічною освітою, що мають практичний досвід застосування методів машинного навчання у вирішенні соціологічних задач або у співпрацюванні з фахівцями з машинного навчання для спільного вирішення певних соціологічних задач. Головними критеріями їх відбору була наявність релевантних даній тематиці наукових праць, наявність вищезгаданого практичного досвіду та наукова зацікавленість у даній темі. Наявність цих критеріїв дозволяє відібрати людей, що мають теоретичне підґрунтя використання машинного навчання у соціології, практичний досвід застосування а також пряму зацікавленість у новітніх зрушеннях щодо цієї теми. Серед даної категорії експертів було опитано:

Тимофія Бріка - випускника Київського Національного Університету імені Тараса Шевченка за фахом “соціологія”, а також випускника декількох західних університетів. Наразі працює дослідником у Київській школі економіки, викладає та має великий досвід у соціологічних дослідженнях, у тому числі з використанням машинного навчання. Сферою наукових інтересів є: кількісні дослідження, науковий аналіз, соціальна стратифікація та мобільність тощо. Має практичний досвід у використанні інструментів для роботи з машинним навчанням, а також співпрацює зі спеціалістами для спільного виконання соціологічних задач.

Дмитра Хуткого - дослідника та практика з інноваційного управління, дослідника українського сервісу, що використовує у роботі машинне навчання “Dozorro”.

Головним завданням в опитуванні цих експертів було дізнатися про парадигмально-методологічну специфіку використання машинного навчання у сучасній українській соціологічній методології, а також зрозуміти для яких соціологічних задач використання машинного навчання буде найбільш

обгрунтованим. На основі виокремленого переліку завдань було віднайдено та опитано наступну категорію експертів.

Друга категорія респондентів - це спеціалісти з машинного навчання, які мають практичний досвід у побудові моделей та алгоритмів для вирішення різноманітних задач (у тому числі і соціологічних). Головними критеріями відбору у даній категорії були: наявність практичного досвіду (більше 3-х років) у роботі з методами машинного навчання. Експертами у цій категоріями стали:

Олег Іванов - завідувач напрямку контент-аналізу у КМІС. Практичний досвід у: текст майнінгу, програмуванні систем збору та обробки текстових даних (text mining, data mining & management, data extraction / ETL), кількісний аналіз текстових даних, у т.ч. із застосуванням машинного навчання, програмування роботів для збору, трансформації та поширення контенту на сайтах соціальних мереж (Facebook, Telegram, Instagram, YouTube, тощо).

Христина Снопик - чотири роки працювала у компанії “Grammarly” на посаді комп’ютерної лінгвістки. Наразі займається виправленням упередженої мови (письмової мови) в англійській, працює над українською «Вікіпедією».

Ганна Пилєва - випускниця магістерської програми Data Science Українського Католицького Університету, має великий досвід в аналізі даних, розпізнаванні образів, обробкою природньої мови та побудовI рекомендаційних систем. Працювала з задачами класифікації, регресії, зниження розмірності та розпізнавання зображень/тексту.

Таблиця 1.2

Перелік респондентів

№	**Експерт**
Категорія 1	
1	Тимофій Брік

2	Дмитро Хуткий
Категорія 2	
3	Олег Іванов
4	Христина Снопик
5	Ганна Пилєва

Для кожної з категорій був розроблений окремий гайд - для першої категорії експертів він був спрямований на виявлення теоретико-методологічної а також парадигмальної специфіки інтегрування машинного навчання в сучасну українську соціологічну методологію. Для другої ж - для виявлення та обґрунтування практичних особливостей застосування того чи іншого методу в українському контексті.

Гайд (як для першої категорії, так і для другої) складався з двох частин - особистого досвіду респондента та перспектив інтегрування у соціологію. Основною відмінністю було формулювання питань та загальна спрямованість опитування.

1.3. Інтегрування методів з інших дисциплін у сучасну соціологічну методологію

Наразі сучасний соціологічний дискурс зосереджений навколо трьох фу ндаментальних питань, які вирішують подальший вектор розвитку методів об робки даних. Такі питання порушені у роботі Волкова, Сугаревського та Тита єва (Волков В. В., Скугаревский Д. А., Титаев К. Д. 2016.), а також у роботах В. Буданова (Буданов, 2007), Касавіна (Касавин, 2004), Волович (Волович В. И., 1974), Паніотто (Паніотто, 1986), Кармінс та Зеллер (Carmines

E.G., Zeller R.A., 1979) тощо.

І перше з питань це: чи потрібно доповнювати сучасну соціологічну мет одологію з інших, суміжних дисциплін задля здобуття кардинально нового ти пу знання?

Значна частина дослідників підтримують ідею міждисциплінарності у соц іології. Головною тезою у підтримку подібного підходу є можливість більш г нучкого дослідження з використанням запозичених з інших дисциплін методі в. Завдяки інтеграції інших методів з суміжних дисциплін, на сьогоднішній день соціологія як наука розширює розуміння про предмет та об'єкт свого дослідження, та має змогу проводити більш змістовний аналіз досліджуваних явищ. Проте варто зазначити, що перелік "суміжних дисциплін" не має чіткого визначення, і відкрита до інтерпретації та обгрунтування.

Заслуговує уваги також думка класика соціологічної науки І. Валлерстайн а про міждисциплінарність та його концепцію «світ - системного аналізу» (Wallerstein I., 1987). На противагу суто міждисциплінарному підходу він про понує монодисциплінарний підхід, у якому всі існуючі дисципліни у рамках однієї сфери не працюють разом, фактично зберігаючи свою автономність, ут ворюють єдину систему. Проте, у підтвердження вже описаного вище феномену вільної інтерпретації поняття "суміжні дисципліни", у ряд з соціологією І. Валлерстайн ставив лише політологію, антропологію, е кономіку та історію. Але зважаючи на сучасний дискурс, є всі основи вважат и, що сюди можна додати і статистику. І. Валлерстайн розглядав соціологію с коріше у суто теоретичному ключі, ігноруючи методологічні завдання, що зазвичай постаю ть перед соціологом.

Але на противагу цьому, деякі з вищезгаданих дослідників

вважають, що популяризація використання міждисциплінарного підходу при зводить до появи у теоретичній парадигмі теоретико - методологічного ред укціонізму. Спрощення основних засад дисциплінарної методології – це спро ба уникнути формальних розбіжностей на межі двох дисциплін, що призводить до примітивізму інтерпретації ключових теоретичних та методологічних засад дисципліни та ставить під питання надійність та валідність результатів дослідження. Тому використання міждисциплінарності має бути зумовлене необхідністю д ослідження об’єкту та предмету на кордоні двох сфер.

Тож, постає логічне **друге запитання**: які є в сучасній методологічній системі значущі проблеми, які мо жуть бути вирішені за допомогою нових, міждисциплінарних методів?

І першою можливою проблемою є гіпотези. Як зазначав Ядов, гіпотезою є :

«*Гіпотеза - головний методологічний інструмент, що організує весь пр оцес дослідження і підпорядковується його внутрішній логіці. У соціологічно му дослідженні гіпотези - обґрунтовані припущення про структуру соціаль них об'єктів, характер зв'язків між досліджуваними соціальними явищами і можливих підходів до вирішення соціальних проблем.»* (Ядов В.А. ст.40)

При формулюванні та підтверженні або спростуванні будь - якої гіпотези дослідник завжди ризикує внести долю суб’єктивності у власне дослідження. Адже навмисно намагаючись підтвердити або спростувати будь - яку з гіпотез, соціолог може потрапити під вплив людського фактору та не об'єктивно оцінити вхідні дані, підлаштовуючи їх під гіпотезу або спеціальни м чином підбираючи дані, які її підтверджують та ігноруючи важливі факти н а користь спростування. Така ситуація зустрічається як на етапі формулювання гіпотез та опрацювання наявної за темою дослідження

літератури, так і на етапі емпіричної частини та аналізу результатів - у формулюваннях анкетних запитань, форматі ведення діалогу з респондентом, інтерпретуванні отриманих показників та компонування їх у певні висновки.

Другим потенційно проблемним місцем звичайного соціологічного дослід ження є валідність (Волович В. , 2010). Це означає, що результати досліджень витримують перевірку будь - я кими іншими методами та демонструють схожий результат, адже вивчають одне й те саме явище і достовірно передають його значення. Якщо два різних методи досліджують одне й те саме явище та на в иході дають певну інформацію про один і той самий предмет, доказом валідн ості таких методів є кореляція між цією інформацією.

Третьою проблемою дуже часто стає надійність даних. По суті надійність – це відсутність випадкових похибок у процесі вимірювання (Alreck P.L., Settle R.B., 1985). Вона перевіряється шляхом перевірки експерименту тим самим методо м, у той самий час та у тих самих обставинах. Якщо результати другого дослі дження співпадають з результатами першого, то його результати є надійними і дослідник може вважати, що у вибраному ним способі дослідження певного соціального явища відсутні похибки вимірювання.

Тож, з урахуванням усього вищезгаданого можна дійти до висновку, що с оціологічні методи мають досить багато можливостей для інтеграції інших методів за умови забезпечення гарних показників якості дослідження.

Тоді постає **останнє запитання**: чим доповнити вже існуючу систему методів, щоб закрити якомога більшу частину (вищезгаданих) проблем сучасної соціологічної методології?

Загалом, зараз у вільному доступі є велика кількість методологічних мате

ріалів з використання машинного
навчання, але цього недостатньо для ствердження про ґрунтовну практику ви користання цих методів в українських соціологічних дослідженнях. Альтерна тивою є велика кількість англомовних матеріалів, проте вони не є адаптованими під український соціально-культурний контекст проведення соціологічних досліджень та використання комп'ютерних технологій для їх проведення. До того ж, як вже було зазначено у попередньому підрозділі, перед дослідниками у такому випадку постає питання пошуку або збору чистого та придатного для аналізу масиву. Компанії - гіганти, такі як Яндекс або Google, (Вайгенд, 2017) світові банки або телекомунікаційні системи володіють величезними м асивами інформації, які не можуть бути оприлюдненими, але використовуют ься у межах самих компаній. Так, аналізуючи поведінку користувачів, вони о рганізовують процес взаємодії з продуктом такми чином, щоб підвищити ріве нь комфорту при користуванні та збільшити кількість продажів. Звичайно, та кі дослідження мають свою комерційну складову, адже підвищивши продажі (за рахунок доречної пропозиції у правильний час та правильній аудиторії) на віть на 0,02%, у перерахунку на грошовий еквівалент це будуть мільйони дол арів. Але такі компанії, що мають доступ до ексабайтів інформації, не мають права її оприлюднювати. Проте незважаючи на це, великі дані досі є дуже пе рспективним матеріалом до соціологічних досліджень. (Cai T., Zhou Y, 2016).

Також важливо окреслити головні проблеми використання машинного навчання як складової дослідження у сфері його соціального контексту. По-перше,
використання великих масивів інформації як нового джерела даних тягне роз уміння зміни соціального простору як такого. Предикативну роль відіграє до повідь компанії IDC (IDC, 2017). У ній компанія наголошує на тому, що у сві

ті простежується стрімке зростання кількості взаємодій індивіда з його гадже том, а це тягне за собою зміну природи соціальної дії та соціальної взаємодії, які необхідно враховувати при дослідженні.

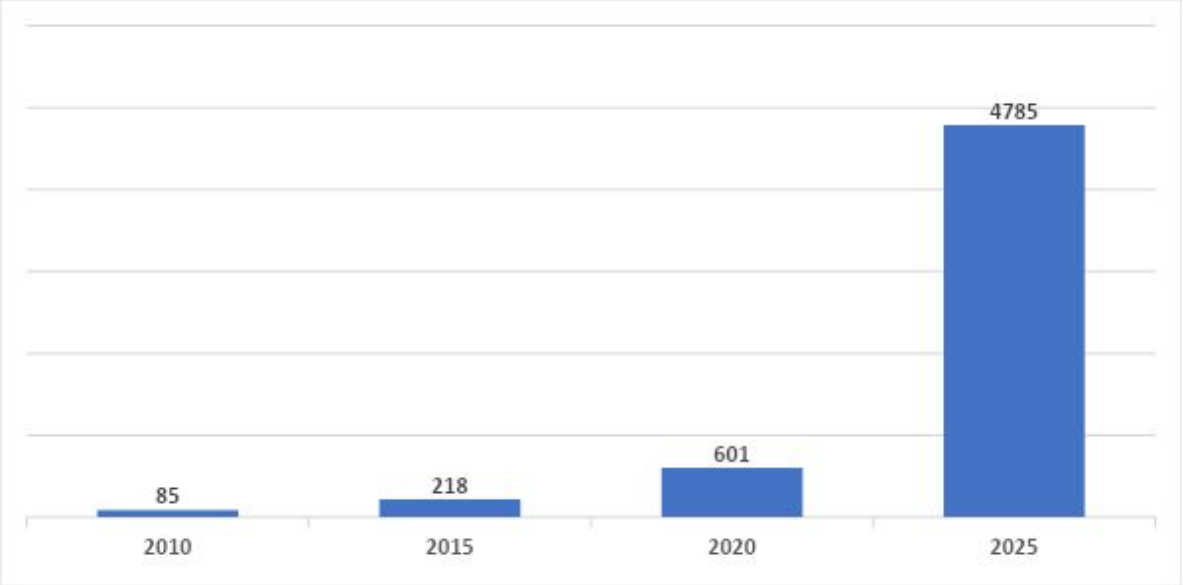

Рис. 1.7. **Кількість взаємодій з гаджетом на день за доповіддю компані ї IDC (2017)**

Друга фундаментальна проблема описана у есе Дональда Кнута ще у 1974 році, і це проблема «Людина - Машина» (Knuth, 1974). У цій роботі Кнут наголошує на відмінностях між актом написання ко мп'ютерної програми та процесом створення предмету мистецтва.

«Наука - це знання, яке ми розуміємо так добре, що можемо навчити й ого комп'ютеру; і якщо ми чогось не зовсім зрозуміли, це мистецтво мати сп раву з цим.» (Knuth, 1974, p. 668).

Дослідник розповідає про аспект створення та сприймання створених ком п'ютером продуктів.

Адже якщо створений продукт або рішення не буде мати аспект тієї самої «л юдяності», то соціум не зможе сприйняти це рішення (якщо йде мова про сит уації з морально - етичним аспектом). Саме через це, за Кнутом, в інтерпрету

ванні даних або у класифікації вхідних значень необхідне втручання дослідни ка.

І останній аспект, на якому наголошував Майкл Барлоу у своїй праці «Rea l Time Big data Analytics: Emerging Architecture» (2013), це проблема застаріл ості даних. Світові банки збирають ексабайти інформації кожного дня, але фа ктична їх обробка не здійснюється. Але тим часом кожного дня та кожного мі сяця ці дані вже перестають бути релевантними, адже в умовах сучасного роз витку суспільства дані можуть вважатися «свіжими» вже дедалі коротший пр оміжок часу. Будь - яка інформація має бути оброблена у найближчий час, а г лобальна система обробки інформації особливо у рамках соціологічних та де мографічних досліджень, має бути автоматизована до найдрібніших деталей (звичайно, виключаючи ретроспективні моделі досліджень).

Проте не дивлячись на ряд проблем соціального значення, використання методів машинного навчання у соціології дає
дослідникам доступ до досить вагомої функції – обробки великих масивів да них задля вивчення соціальних взаємодій.

Якщо говорити про способи обробки даних, то використання методів маш инного
навчання може надати функціонал не тільки для вирішення стандартних зада ч (кластерування, регресія, контент аналіз і т.д), а й привнести у соціологічні методи перспективи нового виду досліджень.

По - перше, завдяки методам машинного навчання ми маємо можливість провести операцію кластерування вхідних даних завдяки різним алгоритмам. Сегментація даних на множини та підмножини здійснюється завдяки самості йно визначеним алгоритмом або дослідником характеристикам. До того ж ви користовується для пошуку нових закономірностей, які до цього не були вио кремлені як значущі. Побічною дією подібних алгоритмів є також виявлення

аномалій, девіантних відповідей респондентів. У сукупності вивчення та вияв лення аномалій та закономірностей у соціальному просторі дає можливість м асштабування досліджень зі студій соціальних взаємодій.

Поряд з кластеризацією та класифікацією, алгоритми машинного навчанн я також закривають завдання регресії – предикативної функції досліджуваних явищ. Наразі це активно використовується у логіці алгоритмічного трейдинг у або на президентських виборах. Даний алгоритм представлений у вигляді й мовірності майбутньої події, що заснована на наївній Байєсівський теорії ймо вірностей (Andrew Gelman, John B. Carlin, Hal S. Stern, Donald B. Rubin, 2003).

Окрім того, різні алгоритми машинного навчання загалом можуть у кінці видавати як бінарне значення, так і множину характеристик, кластери і т.д. С аме тому велика кількість досліджень може бути побудована не тільки на від борі нових, недосліджених характеристик об’єкту дослідження та їх кореляції , а й на відборі релевантного рішення на основі заданих параметрів – тобто за безпечені бінарного вихідного значення.

До того ж, методи машинного навчання (зазвичай нейромережі) використ овують також для обробки цифрових документів. Дані мережі мають можлив ість не тільки розпізнавати будь - який контент, але й самостійно синтезуват и його. Оскільки контент - аналіз є перспективним напрямком соціологічних досліджень та сам контент - аналіз має ряд проблем, пов’язаних зі привнесе нням дослідником суб’єктивності у інтерпретуванні наявних матеріалів, дана тема є перспективною для вивчення.

У глобальній ж перспективі, впровадження методів машинного навчання у сучасну соціологічну методологію є важливим кроком в інтерпрет ації сучасних соціальних взаємодій та підвищенні прогресивності та ефектив ності соціологічних досліджень. Відтак, вони мають можливість мати не тіль

ки номінальну роль у суто академічному середовищі, а здобути практичну пе рспективу у розвитку суспільства.

Світовий банк даних постійно зростає, а динаміка розвитку міждисциплін арних досліджень та впровадження іноватівних технологій у дослідження пе рестає бути поодинокими випадками.

Проте, існує певний соціально - культурний бар'єр, що стоїть на заваді ви користання нових методів у сучасних українських соціологічні дослідження. А саме – недостат ній рівень технічної підготовки сучасних спеціалістів та недостатнє фінансув ання сегменту соціальних досліджень на рівні держави. А це тягне за собою в ідсутність матеріально - технічної бази, програмного забезпечення та мотива ції у співробітників українських центрів соціальних досліджень. У рамках роз гляду цієї теми варто зазначити, що описані в роботі методи вимагають від до слідників певних специфічних знань та навичок міждисциплінарного мислен ня, а також вагомого теоретичного та практичного досвіду роботи з великими даними. Та не зважаючи на те, що базова університетська підготовка в Україні не передбачає соціолога у ролі аналітика великих даних, доповнення власних навичок та їх розвиток доступні кожному досліднику для опанування модифі кованої професії соціолога.

У той же час це породжує поштовх до зміни самої української навчальної системи та перегляду підходу до навчання. Адже впровадження нового, біль ш гнучкого освітнього плану потребує від майбутніх спеціалістів невпинного розвитку міждисциплінарних комунікацій та підлаштування під динамічний розвиток технічних можливостей сучасної наукової парадигми. Таким чином, технологічний прогрес в області досліджень соціального простору породжув атиме новий тип спеціалістів, які матимуть специфічний набір якостей, відмі

нний від тих, що мають дослідники сьогодні.

У сучасному суспільстві досить складно відслідкувати та окреслити фундаментальні зміни та модифікації, які породжує технологічний прогрес. Проте ми можемо дати приблизну оцінку наявним методам у соціологічних дослідженнях та у методології машинного навчання, аби сформувати перспективи їх інтеграції в сьогоднішню парадигму та окреслити нове поле діяльності у сучасному науковому доробку.

У такій ситуації варто згадати про теорію аномії Еміля Дюркгайма. Згідно до неї, коли суспільство знаходиться на хиткому проміжному етапі та в умовах суцільної несформованості нових цінностей та відмирання старих, суспільство поринає у стан специфічної невизначеності та кризи, що і названа аномією:

«Одна і та ж криза може сколихнути існування індивіда, порушити рівновагу між ним і його середовищем і в той же самий час звернути його альтруїстичні нахили в стан, що збуджує в ньому думку про самогубство.» (Дюркгейм, ст.20, 1897)

Схожий феномен можна спостерігати і в академічному полі, коли старі методи обробки даних вже застаріли та стали неефективними в умовах зміненого суспільства, а рівень технічної підготовки спеціалістів ще недостатній, аби переходити на нові, які використовуються у більш технологічно розвинених суспільствах. Тому дуже важливе своєчасне впровадження нових методів в усталену методологію, аби цього розриву та кризи не відбувалось, а навчання та зміни проходили планомірно і без впадання у шоковий стан.

Якщо ж говорити про перспективи інтегрування методів машинного у сучасну соціологічну методологію, необхідно згадати і про практичне застосування цього методу у глобальній перспективі.

Наприклад, вище було зазначено, що метод машинного навчання з підкрі пленням не цікавий для розгляду соціологами як такий, але є випадки, коли ві н латентно також може приносити багато користі для майбутніх дослідників. Наприклад, з перспективою використання безпілотних автомобілів пов'язани й досить вагомий факт, а саме – постійна обробка величезної кількості інфор мації на дорозі. Якщо розглянути дану перспективу у найближчому майбутнь ому то можна дійти до висновку, що кількість даних, які обробляє один автом обіль впродовж дня є цінним джерелом знань про соціальні взаємодії. Аналіз таких даних (з використанням тих самих методів машинного навчання) є кл ючем у вирішенні багатьох соціальних, економічних, демографічних, а також екологічних проблем. Опади у різний час у різних районах, специфіка повод ження на дорозі велосипедистів, забрудненість вулиць або безпека дітей на д орозі - все це може бути предметом вивчення соціологів. Цей приклад є дем онстративним, але поки не втіленим у життя, але наразі ми маємо велику кіль кість прикладів з подібним механізмом дії. Генерування великих масивів нео броблених даних ставить перед дослідниками завдання термінового пошуку методів обробки, впорядкування та аналізу цих даних та прийняття конкретн их рішень на основі результатів цих досліджень про подальший розвиток сус пільства, чого неможливо зробити з використанням локальних досліджень.

Інший приклад – генетика. Наразі практика здавати тест на генетичні схил ьності набула розголосу у західних країнах (наприклад, Великобританії). Завдяки цьому у вчених з'явилося нове джерело великих да них для демографічних досліджень та суттєвих соціальних зрушень. Наприкл ад, у 2019 році, міністр охорони здоров'я Великобританії виступив з заявою, що у найближчий час ДНК - тест новонароджених дітей стане рутиною для б ританців. (All children to receive whole genome sequencing at birth, under ambitions laid out by Matt Hancock, 2019). За його словами, це допоможе уникнути у подальшому серцево - судин

них захворювань та розвитку онкологічних захворювань. Якщо співвіднести поглиблені демографічні характеристики кожного індивіда, співставити їх з с оціальним бекграундом (адже більшість людей, що пройшли подібний тест Д НК, мають спеціальні програми для контролю своїх схильностей), то можна о тримати досить розповсюджений, всеохопний та вичерпний матеріал для дос ліджень всередині кожної країни та загалом.

Отже, з огляду на всі вищезгадані аспекти можна дійти до висновку, що ін тегрування методів машинного
навчання у сучасну соціологічну методологію є перспективним та може суттє во розширити дослідницьке поле соціологічного дослідження.

РОЗДІЛ 2. ПРАКТИЧНЕ ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ВИРІШЕННЯ СОЦІОЛОГІЧНИХ ЗАВДАНЬ

2.1. Перелік методів машинного навчання, які можливо інтегрувати у соціологію

Даний розділ передбачає розкриття відповідей першої категорії експертів, які є соціологами з практичним досвідом використання машинного навчання для вирішення соціологічних задач. Ключовим у опитуванні було дізнатися чи є перспективним інтегрування машинного навчання у сучасну соціологічну методологію. Варто зазначити, що отримані відповіді мали більш-менш одностайний характер та можуть бути інтерпретовані як такі, що мають ствердну відповідь на це запитання. Загалом експерти висловлювали позицію, що інтегрування машинного навчання у соціологічну парадигму не

тільки має місце в українському контексті, але й є парадигмальною необхідністю, що зумовлена більше не потребою в новому інструментарії, а у зміною у дисциплінарній площині:

“У мене досить радикальна позиція: соціологи або почнуть використовувати весь спектр сучасних кількісних методів, або соціологія як наука помре і перетвориться на «науковий комунізм» — псевдонауку, яка обслуговує ідеологічно ангажовані групи інтересів.” (Олег, 3)

Проте це не є причиною нівелювати важливість впровадження нових методів у методологічний інструментарій дослідників-соціологів. Тут необхідно зазначити, що дана інтеграція не атакує значущість та теоретичну обумовленість наявних в українській соціології методів, а лише впроваджує додаткову можливість дослідження там, де з тих чи інших причин класичні методи збору даних або їх обробки вже не працюють. Нижче наведена думка одного з респондентів, що ілюструє цей процес:

“В усьому світі опитувальні та інші контактні методи соціологічних досліджень переживають кризу. Неухильно падає частка відповідей (response rate). У західних країнах F2F майже «померли», в Україні цей процес йде у тому ж напрямку і він неухильний. Телефонні опитування та опитування через Інтернет, нмд, є лише тимчасовим рішенням. В умовах інформаційного перевантаження, яке дедалі зростає, йде боротьба за увагу людини.”(Олег, 3)

Також варто зазначити про таку дисциплінарну специфіку подібної інтеграції. У ході дослідження було висунуто припущення, що однією з проблем такої інтеграції (що пов’язане зі згадуваним у попередніх підрозділах явищем методологічного редукціонізму) є неузгодженість підходу соціологів та спеціалістів з машинного навчання до побудови

дослідження. Адже класичне соціологічне дослідження виходить з теоретичного базису, на основі якого створюється гіпотеза, а у машинному навчанні радше виходять з практичного вивчення наявних даних, і формування проблематики дослідження з огляду на них. І один з респондентів на початку інтерв'ю підтвердив цю гіпотезу.

"Справа в тому, що люди що працюють з машинним навчанням, для них дуже часто не важлива тема дослідження, для них більше важливо що є алгоритм, і цей алгоритм є цікавим самим по собі і хочеться його застосувати будь-де". (Тимофій, 1)

Тобто, формулюючи тему та мету свого дослідження, соціолог з більшою ймовірністю спирається на теоретичного підґрунтя, аніж з незрозумілих, аномальних, девіантних, або просто цікавих даних. Проте це припущення було спростоване експертами. Головною тезою на противагу цього припущення було те, що масиви даних можуть стимулювати теоретичне мислення і навпаки. Обидва варіанти, подібно до методів дедукції та індукції, застосовані в обидвох дисциплінах, і ніякого практичного обмеження у даному факті немає. Той же експерт, що і з попереднього цитування, доповнив свої слова більш розлогою аргументацію, що спростувало перше розуміння попереднього тезису:

"Навіщо я роблю цей вступ? Щоб атакувати стереотип, що нібито в природничих науках всі такі орієнтуються на дані, а ми, соціологи, орієнтуємося на теорію". (Тимофій, 1)

Проте варто зазначити, що один з експертів наголошував на тому, що все ж даний факт присутній в українському академічному та освітньому середовищах. У цьому контексті варто розглянути освітні заклади як агентів конструювання сприйняття дискурсу дослідження молодими експертами. Формуючи чіткі рамки про дослідницький процес, це унеможливлює

гнучкість дослідження та гальмує впровадження нових методів та підходів до досліджуваних явищ.

Варто сказати також про роль дослідника та необхідного набору навичок для роботи з машинним навчанням. Декілька експертів зазначали на двох опціональних рівнях, на яких може розвиватися дослідник у ході вивчення машинного навчання. Перший- це так званий "data analyst" (аналітик великих даних), який знає основні принципи машинного навчання, добре розбирається у способах його використання, може самостійно виконувати певні універсальні задачі за допомогою готових інструментів або грамотно поставити завдання спеціалісту з машинного навчання, якщо потрібно розробити більш кастомізовану модель.

І другий рівень - це дослідник, що повністю заглибився у машинне навчання, має навички написання алгоритмів на Python та може самостійно створювати унікальні алгоритми для кожного завдання.

Також у ході опитування, експерти виокремили декілька соціологічних задач, для вирішення яких доцільно впроваджувати машинне навчання у сучасну соціологічну методологію.

Перший з них - це завдання класифікації.

"В принципі в соціальних науках де найчастіше застосовують методи машинного навчання? Там, де вони працюють найкраще. Вони працюють найкраще в класифікації. Розпізнати щось і сказати що щось належить до групи 1 чи групи 2. Призначити нулик або одиничку." (Тимофій, 1)

Друге з завдань, яке можна вирішити за допомогою машинного навчання - це контент аналіз та мережевий аналіз. У основі цих методів також лежить класифікація, проте принцип дії та вихідний результат абсолютно відрізняються. Класичний контент-аналіз широко використовується

дослідниками та має широкий набір тематик для вивчення.

“Такі самі алгоритми можна застосувати для аналізу текстів, наприклад. Чи належить текст до якоїсь групи (...)?” (Тимофій, 1)

У сучасній соціологічній парадигмі контент-аналіз розглядається як метод кількісно-якісного дослідження для вивчення тенденцій у соціально-культурному просторі шляхом аналізу будь-яких документів та соціального контексту, у якому вони були створені. (Корсунська, Семенова, 2010). У рамках даного методу розмежовують декілька напрямків.

Напрямок ідентифікації передбачає аналіз досліджуваних у документі явищ з огляду на реальний контекст подій. Це є індикатором пріоритетності певних подій, ознак або явищ для автора документу або всього соціального середовища. Морфологічний напрямок же спрямований на вивчення орфоепічних, синтаксичних, лексичних або граматичних характеристик мови певного документу та виокремлення певних ознак або правил, які є значущими для охарактеризування спільноти або явища.

І останній напрямок це ефективний - це аналіз характеристик, за якими певний документ стимулює емоційну навантаженість, характеризування інтенсивності та результату, на який спрямовані ці характеристики.

Проте для аналізу великої кількості документів потрібно залучити велику кількість людських ресурсів, що буває невиправданим, адже обробка такої кількості даних вручну може суттєво вплинути на результати цієї обробки. У такому випадку доречно використовувати методи машинного навчання.

Ось один з можливих кейсів:

“І ось вони робили текстовий аналіз книжок, які видавалися в Німеччині в XVII-XVIII ст. За допомогою методів машинного навчання вони змогли просто за назвами цих книжок класифікувати, чи буде ця книжка скоріше

антисемітською чи ні." (Тимофій, 1)

Окремо потрібно окреслити інтеграцію методів машинного навчання у дослідженнях соціальних мереж. Це доволі розповсюджений метод контент-аналізу, адже соціальні мережі дозволяють відслідковувати динамічні реакції певної спільноти на окремі явища або події. Проте забезпечення репрезентативності вибірки в умовах обмеження людськими ресурсами є проблематичним.

Проте у випадку застосування машинного навчання, це завдання можна вирішити без таких витрат на операційну діяльність дослідників, та забезпечити більшу репрезентативність вибірки завдяки більшим об'ємам даних. До того ж, дані з соціальних мереж є відкритими та відносно простими в здобутті.

"Другий приклад - це, власне, мережевий аналіз тому, що мережі це, власне, дуже універсальний феномен. З мережами працюють всі, з мережами працюють фізики, біологи, геологи, генетики, соціологи, мережі можуть бути зовсім різні." (Тимофій, 1)

Демонстративним прикладом контент-аналізу за допомогою машинного навчання може бути, наприклад, аналіз публікацій-ретвітів у Twitter, завдяки якому соціологи можуть простежити за динамічними змінами у соціальній взаємодії у таких специфічних осередках цієї взаємодії. Показовим у реакції подібних інформаційних збудників є дослідження ролей та статусів залучених в обговорення індивідів, а також їхня сегрегація на основі займаної позиції. Наприклад, у разі вивчення рівня тривожності або стресу у певної групи індивідів (зважаючи на їхнє місцезнаходження, вік, стать, тощо), соціолог може встановити алгоритм з набором порогових значень для різних ознак взаємодії індивіда з іншими особами (як вживання специфічної

лексики, обсяг твіту, загальне емоційне забарвлення тощо).

Проте варто зазначити, що лонгітюдні дослідження в мережевому аналізі - це складне питання, і про грунтовну практику вирішення цього завдання методами машинного навчання ще говорити зарано.

“Саме складне питання - це лонгітюдні. Це розвиток мережі з часом”. (Тимофій, 1)

Завдання лонгітюдних досліджень розкриває проблему каузальності - одного з важливих напрямків дослідження сучасними соціологами, проте чого не можуть надати методи машинного навчання. Але варто зауважити, що це питання вже розглядається спеціалістами з машинного навчання і є вірогідність, що практика використання машинного навчання для ілюстрації каузальності найближчим часом увійде у сучасну науку. Ось один з прикладів цієї проблеми:

“Сучасні алгоритми машинного навчання вони можуть вам ідеально розказати чи є у вас рак шкіри чи нема, але скоріш за все вони не зможуть вам сказати чому. І люди, які працюють з машинним навчанням називають це “black box”, коли щось відбувається за лаштунками алгоритму, що нам не до кінця зрозуміло” (Тимофій, 1)

Наступним завданням, у вирішенні якого може бути задіяне машинне навчання, це логістична регресія.

Регресія також є фундаментальним поняттям у машинному навчанні. Проте, на відміну від класифікації, які є базовим завданням машинного навчання, логістична регресія є його методологічним принципом роботи. Регресійний аналіз - це статистичний метод аналізу залежності між залежною змінною Y та незалежним предиктором X (Дрейпер, Сміт, 2007).

У ході застосування даного методу можна розглядати також застосування

множинного регресійного аналізу, у ході якого встановлюється залежність між залежною змінною Y та певною множиною незалежних змінних X1...Xn. Поширеною є практика використання даного методу для виокремлення значущих факторів, які діють на залежну змінну.

"Якщо ви під час своєї університетської освіти вивчили що таке логістична регресія і як її можна інтерпретувати (а саме краще щоб ви щось там і від руки робили, (...)) я вас запевняю, цього буде достатньо щоб потім переключитись на машинне навчання (...)Тому що машинне навчання це ду-у-у-же складна логістична регресія. Якщо ви зрозумієте дуже просту логістичну регресію, то розуму щоб осягнути дуже просту вам точно вистачить."(Тимофій, 1)

Якщо узагальнити, то регресійний аналіз закриває два фундаментальні завдання. Першим є предикативний аналіз, тобто прогнозування позиції залежної змінної у випадку впливу на неї однієї або декількох незалежних змінних. У ході такого дослідження діяльність соціолога стає ітераційним процесом, у ході якого дослідник виокремлює набір змінних, які виступатимуть якісними тригерами передбачення. Друге завдання - це стандартна функція регресійного аналізу, тобто дослідження якою мірою незалежна змінна пояснює залежну.

У соціологічних дослідженнях такий метод зазвичай застосовують щоб пов'язати набір явищ у соціальному просторі та для вивчення міри впливу одного явища на інше. Для прикладу можна навести ряд досліджень, такі як встановлення впливу рівня заробітної плати у різних категоріях населення на рівень міграції у цих категоріях, моделювання показників кількості електоракту певного політичного блоку в залежності від обсягу страт тощо.

Також завдяки відповідям експертів було виокремлено декілька важливих проблем (як практичних так і дисциплінарних, дискурсивних та

парадигмальних), які складають основні труднощі в використанні машинного навчання для вирішення соціологічних задач в Україні. І першою з проблем можна назвати проблему збору даних. Як вже було зазначено раніше, процес збору великих даних, які використовувалися раніше в Україні (інтернет- або телефонні опитування) потроху втрачають свою ефективність. Проте це ілюструє лише те, що дані оновлюються та переходять у інший формат, а отже і методи їх збору мають відрізнятися від попередніх:

"Водночас, зростає обсяг «неспровокованих» персональних даних: записів у соцмережах, історій перегляду сайтів («кукі»), історій пересування (GPS трекінг), записів з камер спостереження, тощо. Це все вже аналізується. (...). Всі ці дані є «великими» і можуть ефективно бути проаналізовані лише з допомогою сучасних кількісних методів, у т.ч. машинного навчання." (Олег, 3)

Тут також варто згадати і перешкоджання вільному збору інформації з огляду на збереження особистого простору користувача, політики конфіденційності і т.д. Наприклад, е 2016 році в Official Journal of the European Union були видані регулятивні вимоги до захисту даних, що були викладені у документі General Data Protection Regulation (The European Parliament and the Council of the European Union, 2016) . В ньому викладені основні правила обробки персональних даних, які розповсюджуються на всіх країн-членів Європейського союзу. Серед положень є, наприклад, мінімізація зібраних даних, тобто збір тих даних, які мінімально закривають потреби компанії і заборона збирання даних у більшому обсязі, ніж це необхідно для першопочаткової цілі, або обмеження терміну збереження інформації у збирача.

"Є світовий тренд до зменшення можливостей «легального» збору даних для КА: GDPR, TOSи веб-сервісів, які забороняють автоматизований збір

даних, урізання можливостей відкритих API. Всі це повсюдно порушують, але з іншого боку компанії, які накопичують ці дані (такі як Facebook та Google) всіляко намагаються технічно перешкодити їх отриманню дослідниками. Це «гонка озброєнь», яка в окремих випадках може вийти і в легальну площину з мільйонними позовами." (Олег, 3)

Другою проблемою варто зазначити професійну та математичну підготовку дослідників. Професійна, або технічна підготовка соціологів впливає на їхню можливість швидко та без особливих труднощів розібратися у професійній літературі. З огляду на швидкий темп розвитку індустрії машинного навчання, професійна література не встигає з'являтися у достатній кількості, і люди, що цікавляться подібною темою мають мати навички пошуку цієї інформації з різноманітних джерел. До того ж, з огляду на складність літератури, для її розуміння потрібно мати певну гнучкість міждисциплінарного сприйняття інформації та набір специфічних знань з цієї тематики. До того ж, для розвитку у подібній галузі потрібно мати і певні специфічні навички (такі як програмування за допомогою Python), які зазвичай не передбачаються для вивченні на спеціальності "соціологія" у вищих навчальних закладах України.

"На сьогодні володіння, принаймні, однією мовою програмування є такою ж вимогою до професійної придатності на глобальному ринку праці, як і володіння англійською мовою. Застосування машинного навчання без написання коду можливе лише у готових програмних пакетах, які швидко застарівають та не забезпечують достатній рівень гнучкості для всіх дослідницьких ситуацій."(Олег, 3)

До того ж, важливою проблемою є також недостатнє фінансування та затребуваності на українському ринку. Наразі спостерігається тенденція збільшення фінансування маркетингових досліджень приватними

компаніями, а отже і розвиток цієї галузі в цілому. З огляду на те, що дослідження соціального простору як такого напряму не приносять прибуток для бізнесу, вони не стоять у пріоритеті в фінансуванні як на приватному, так і на державному рівнях.

“Все ще сировинний характер української економіки. Інтелектуальний продукт не має такого рівня попиту на українському внутрішньому ринку, як на західних. Наш славетний IT-сектор — це на 90% аутсорс і аутстаф для західних компаній. На жаргоні політтехнологів, піарників і т.д. «зробити соціологію» значить провести опитування. Ні з якими іншими методами (тим більше неконтактними) соціологічні дослідження у широкому вжитку не асоціюються.” (Олег, 3)

І останньою проблемою, яку виділили експерти у відповідях - це дисциплінарно-дискурсивне питання заангажованості сучасної соціології. Більшою мірою вона перегукується з попереднім пунктом, адже свідчить про певний рівень неспроможності сучасного українського контексту використовувати більш прогресивні методи та відсутність попиту на дослідження, що не мають на меті певної фінансової прибутковості у разі винайдення шляхів впливу на соціальне середовище або кращого його вивчення та формування відповідних характеристик продукту.

“Сучасна соціологічна наука (у світі загалом) є ідеологічно заангажованою. Більшість наукових грантів мають «ціннісне» навантаження. Методологічні розробки з кількісних методів не вважаються вартими фінансової підтримки, бо їх «профінансує індустрія». Однак, у випадку МН на практиці це не соціологічна «індустрія», а IT, куди і йдуть соціологи з відповідними знаннями і вміннями.” (Олег, 3)

До того ж необхідно згадати перспективи інтегрування методів машинного навчання в українські державні сервіси. Один з експертів є

дослідником подібного сервісу “Dozorro” який виконує функцію громадського контролю за державними закупівлями на платформі “ProZorro”. Даний вид компіляції поєднує у собі як соціальну взаємодію та можливість її вивчення, так і методи машинного навчання, що використовуються у сервісі. До того ж, даний ресурс може розглядатися як джерело великих даних, які збираються та систематизуються.

“Було оголошено що така методика застосовується, наприклад, в “Dozorro”. Головний архітектор (програміст) який замислював, реалізовував та контролював кодування всієї конструкції “Dozorro”. Він заявив, що дійсно там використовується елементи машинного навчання, тобто це штучний інтелект який в тому числі працює з елементами машинного навчання”. (Дмитро, 2)

За словами експерта, даний сервіс працює з ризик-індикаторами, що не дозволяють проводити незаконні маніпуляції з державними закупівлями. Наприклад, розбивати закупівлю на окремі складові та проводити їх як менші закупівлі для надання їм статусу “допорогових”, адже у випадку “надпорогових” службовці мають більше вимог до звітування. І саме щоб уникнути подібних махінацій, використовується штучний інтелект для виявлення сумнівних дій.

“Суть в тому, що він працює з ризик-індикаторами публічних закупівель. І він звіряє реальні дані які поступають від закупівель (підприємства, державні агенції, комунальні підприємства) і цей штучний інтелект здійснює автоматичну перевірку даних про ці закупівлі з списком ризик-індикаторів”. (Дмитро, 2)

У контексті України дані сервіси є перспективними у розрізі налагодження ставлення суспільства до роботи державних та

адміністративних органів та динамічного відслідковування цього ставлення.

Отже, у цьому підрозділі були визначені основні сфери застосування методів машинного навчання для вирішення соціологічних задач, а саме: задача класифікації, кластеризації, регресійного аналізу та контент-аналізу. А також були визначені основні проблеми, які можуть стати на заваді інтегрування методів машинного навчання у сучасну соціологічну методологію.

2.2. Використання алгоритмів машинного навчання для завдань класифікації та кластеризації

Розглядаючи питання використання алгоритмів машинного навчання для завдань класифікації у соціологічному дослідженні перш за все варто зауважити, що завдання класифікації є одним з базових завдань сучасного машинного навчання, від якого історично сформувалися й інші методи. І частково завдання класифікації присутні у всіх методах, що описані у наступних підрозділах, і саме тому воно наведене першим підрозділом з аналізу інтерв'ю з експертами.

Проте, даний розділ поділяється на аналіз завдань класифікації та кластеризації, адже обидва методи є схожі за дією, проте мають одну відмінну рису: при процесі класифікації дослідник обов'язково задає набір критеріїв за якими дані розбиваються на групи, а кластеризація групує дані без попереднього окреслення критеріїв.

Перш за все, необхідно зазначити перелік досліджень, у яких використовується класифікація. Частково цей перелік перегукується з методами, що вже були зазначені у попередніх методах, адже класифікація у тому чи іншому вигляді- це стандартна функція машинного навчання.

І перш за все, це можливості сегментації населення на певні групи та перерахунок об’ємів даних груп. Варто зазначити, що подібна функція має досить обмежену дію і саме тому носить для соціології досить номінальний інтерес. Подібні задачі зазвичай представляють найбільший інтерес в області медичного діагностування:

“Самий популярний приклад - це розпізнавання, чи є у людини рак шкіри. Ось, методи машинного навчання можуть вам сказати - чи є у вас рак шкіри чи нема, нолик або одиничка” (Тимофій, 1)

Другою важливою сферою застосування класифікації є робота з документами - це класифікація текстів за наповненням, тематикою або будь-яким іншим синтаксичним критерієм. Детальніше цю сферу застосування буде розглянуто у підрозділі 2.4. Варто лише зазначити, що у цей перелік входить також і аналіз тональності тексту, що дозволяє зробити глибший аналіз усього контенту.

Ну і останньою сферою застосування є пошук аномалій, який може бути застосований у будь-яких варіаціях і для різних задач дослідження. Для кращого розуміння нижче наведено схему роботи класичної класифікації.

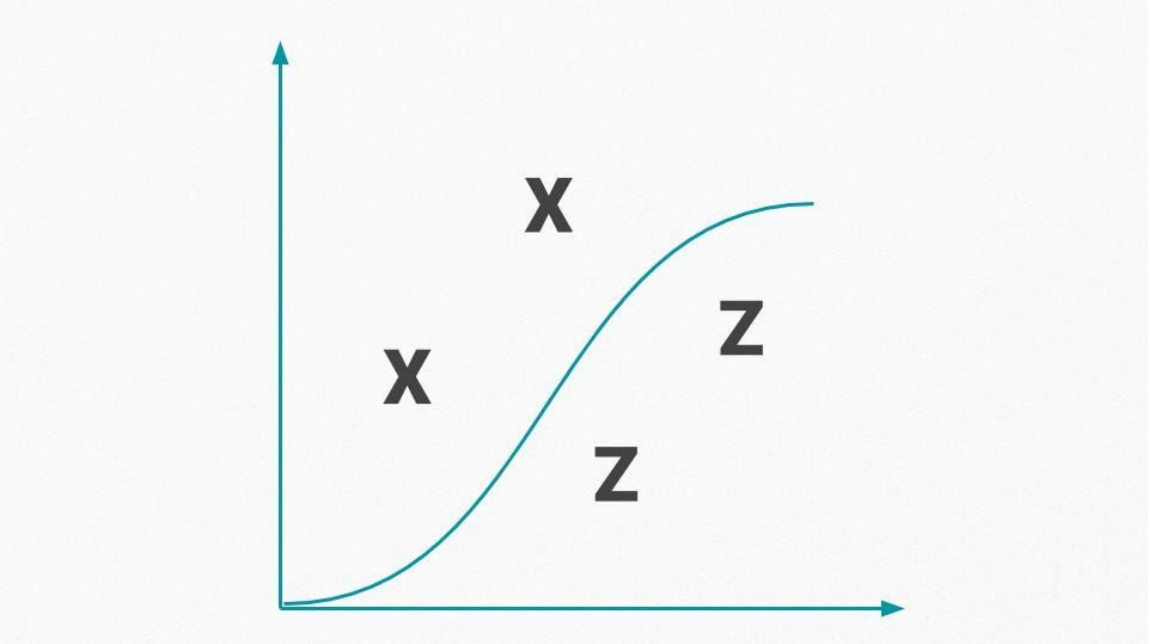

Рис. 2.1 **Схема роботи класичної класифікації**

Якщо говорити про задачу кластеризації, то тут можна спостерігати вже більше перспектив. Наприклад, задача виокремлення характерних рис портрету індивіда, пошук його специфічних особливостей та вподобань. Ось як застосовувала завдання кластеризації одна з експерток (варто зауважити, що тут більше маркетингова цінність завдання, проте даний досвід допустимо екстраполювати на соціологічні задачі):

“Також займалася класичними задачами аля “кластеризація” для того, щоб зрозуміти які є клієнти в фінального клієнта і яким чином ми можемо взаємодіяти з цими клієнтами для того, щоб зробити якусь стратегію того, як до кожного клієнта більш детально підійти і врахувати його особливості взаємодії з нашим продуктом” (Ганна, 5)

Якщо говорити про соціолого-методологічну практику, то найпоширенішим прикладом кластеризації є кластерний аналіз.

Метод кластерного аналізу - це метод, завдяки якому множини впорядковуються у кластери на основі певних характеристик. Так, кожен кластер вміщує у себе групу схожих за характеристиками елементів, проте

кожен кластер окремо суттєво відрізняється від іншого. (Айвазян, Бухштабер, Енюков, Мешалкін, 1989). Даний аналіз доцільно проводити коли дослідник ще не володіє характеристиками досліджуваного явища, не вивчений його вплив на дані та їх кількість занадто велика для аналізування всієї сукупності одразу.

У соціології ж цей аналіз проводиться з використанням методу k-середніх;

“Для кластеризації використовувала k-means, наприклад кластеризацію в кілька способів: в кількох задачах робила, там, знайти дублікати описів товарів. Там треба було якимось чином токенізувати або розбити на маленькі частини текст і потім за набором слів ми визначаємо ці фрази достатньо схожі чи не схожі.” (Ганна, 5)

Метод k-means - один з розповсюджених методів кластерного аналізу, який полягає у зменшенні сумарного квадратичного відхилення точок кластерів від центрів цих кластерів. Або якщо стисло, то мінімізує відстань точки до центру кластеру.

Подібний результат досягається у ході виконання певного алгоритму дій. Перш за все, соціолог має самостійно виокремити кількість кластерів на основі спостережень очевидних сукупностей та власного досвіду. Після цього алгоритм, зважаючи на задану кількість кластерів, формує випадковий відбір їх центрі. Усі дані групуються навколо згаданих центрів шляхом визначення належності до найближчого центру. І тільки після цього всі дії ітеруються для забезпечення сталості кожного з центрів і повторюються до того моменту, доки не будуть встановлені оптимальні відстані для кожної множини.

Проте, машинне навчання вносить нові можливості у даний метод (Kohonen, Somervuo, 1998). Адже в машинному навчанні також є метод k середніх, просто принцип дії трохи інший.

Даний метод має просте втілення та високу якість результатів, за що і використовується дослідниками часто. В основі цього методу лежить один найголовніший параметр - значення k, тобто кількість кластерів, на які необхідно розподілити дані.

Отже, як висновок можна сказати, що завдання класифікації та кластеризації у соціологічному досліджені можна виконати за допомогою методу машинного навчання, а також даний метод лежить у основі всього машинного навчання, що робить доцільним його вивчення на початку ознайомлення з методами штучного інтелекту.

2.3. Використання алгоритмів машинного навчання для завдань регресійного аналізу

Завдання регресії в соціологічних дослідження є затребуваним як в академічному, так і практичному середовищі. Як вже було зазначено у попередніх підрозділах, регресійний аналіз у соціології - це статистична модель аналізу залежності залежної змінної від незалежної. Така модель сама по собі забезпечує дві функції - це функція передбачення, у ході якої соціолог виокремлює якісні тригери передбачення за умови впливу на залежну змінну декількох незалежних.

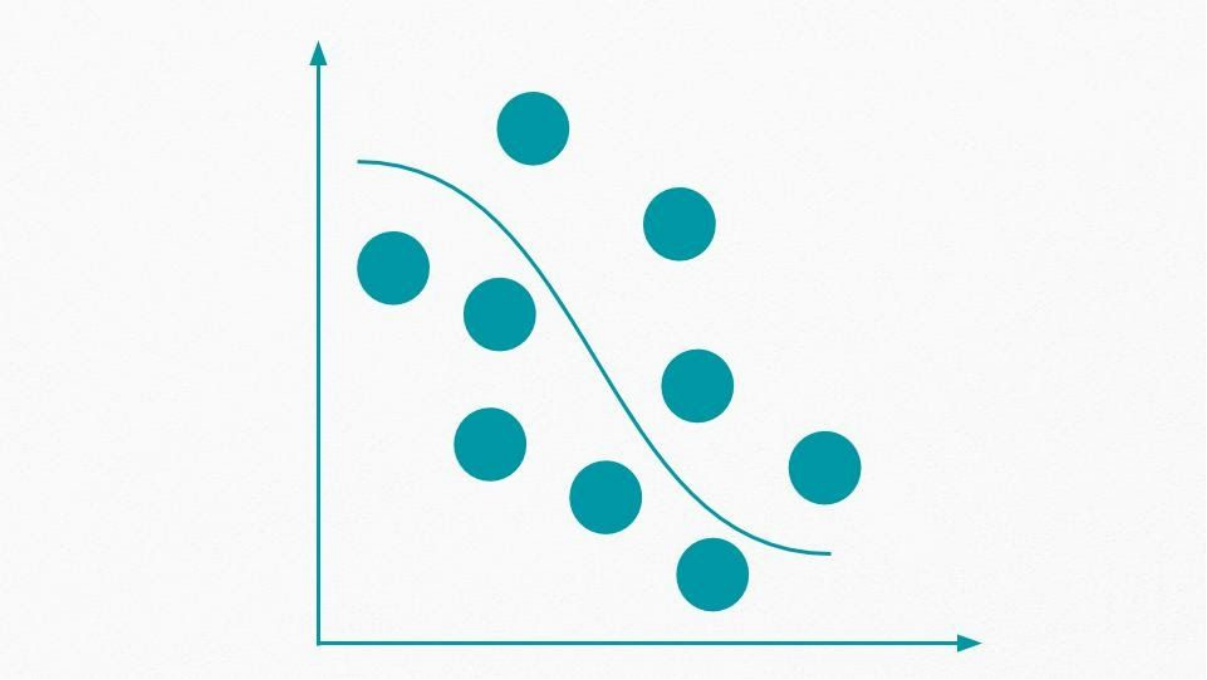

Рис. 2.2 **Схема роботи класичної регресії**

Проте існують також і інші методи регресійного аналізу, що включають у себе методи машинного навчання. Всі вони також мають на меті проведення лінії на графіку, яка відображає залежність змінних.

“Наприклад часові продажі сьогодні і часові продажі, які були вчора. Бо по факту продажі сьогодні вони залежать від того, що було історично, або від того, що було певний період тому. Наприклад, продажі в сьогоднішній понеділок схожі на продажі в попередній понеділок, але не схожі на продажі в неділю, тому що в неділю ми маємо іншу тенденцію” (Ганна, 5)

Варто зазначити, що регресія буває різних видів, наприклад, дві основних - це лінійна та поліноміальна. Лінійна регресія зображується як пряма лінія, поліноміальна - кривою лінією.

Лінійна регресія також може бути двох видів - проста (одновимірна) та багатовимірна. Одномірна регресія - це функція, яка ілюструє модель залежності між незалежною вхідною змінною та залежною вихідною. Багатовимірна регресія ілюструє залежність між декількома незалежними

вхідними змінними та однією вихідною залежною змінною. Проте вона все ж залишається лінійною регресією, адже передбачає лінійну комбінацію вхідних значень.

Якщо ж говорити про поліноміальну регресію, то це вже нелінійна комбінація вхідних значень, тобто серед цих даних можна знайти і більш експоненціальні величини (синуси, косинуси і тд). Перевагою цього методу можна вважати швидкість створення моделі (як з використанням машинного навчання так і без) та інтуїтивна зрозумілість цього методу, тобто вона може бути застосована у ширшому спектрі сфер та великою кількістю спеціалістів. Проте варто зазначити, що при використанні нелінійних даних, можуть виникнути певні складнощі в моделюванні, та з великими об'ємами даних вона може бути не так ефективна.

Проте є ще один вид регресії, який часто заплутує молодих фахівців, що починають розбиратися з теорією машинного навчання. Цей вид-логістична регресія, і фактично вона знаходиться ближче до завдань класифікації, аніж регресії. Подібний феномен пояснюється (а разом з тим і підтверджує його) тим, що більшість сучасних класифікаторів при більш детальній розробці можна перетворити у регресію, і існують випадки, коли ця межа може бути непомітною. Фактично логістична регресія - це лінійний класифікатор, що по суті являється класифікатором, проте має якісну здатність до прогнозування ймовірності.

Проте все вищесказане відноситься більше до базового функціоналу машинного навчання та перекриває невелику кількість задач. Проте за даними відповідей експертки, з більш уніфікованими завданнями краще справляються нейронні мережі:

"Але зараз все ж більше використовують нейронні мережі для аналізу sequence, для аналізу послідовностей, та сама ідея як з текстами, просто ми

кодуємо послідовність або беремо їх певними пачками і подаємо в алгоритм і запускаємо його на передбачення." (Ганна, 5)

Нейронні зв'язки створені з послідовно сполучених або взаємопов'язаних груп нейронів. Усі вхідні дані надходять до нейронів у вигляді лінійної комбінації з великою кількістю змінних. Кожна функціональна змінна має свою вагу, тобто значення, на яке множиться дана зміна. Але потім модель змінюється та приймає форму нелінійної, щоб забезпечити можливість відслідковувати складні нелінійні зв'язки. Найчастіше за все нейронні мережі є багатошаровими, і від вхідного шару дані поступово подаються на наступні.

Головною перевагою подібного методу є побудова складних, багатошарових, нелінійних взаємозв'язків. Також варто згадати про здатність аналізувати навіть неструктуровані дані, що значно полегшує вивчення змінних. Ну і наостанок ще одна перевага - це постійне покращення моделей за рахунок постійного навчання на нових даних.

Проте варто зазначити також і негативні сторони застосування нейромереж для регресійного аналізу. По-перше ця модель є складною, а отже її складніше зрозуміти, розробити та застосувати. Тому тут знадобиться втручання професійних фахівців з машинного навчання. По-друге використання нейромереж не є швидким, адже їм потрібен час на навчання, а дослідникам - на налаштування параметрів. І по-третє нейромережі не підходять для застосування на малих об'ємах інформації, тому що для навчання їм потрібно дуже багато даних.

Якщо говорити про завдання регресії потрібно згадати також використання дерев рішень або, як їх ще називають, "випадкового лісу" (random forest). Загалом, дерево рішень - це набір правил, за якими ієрархізована система надає кожній змінній свій вузол, який приймає рішення щодо цієї змінної. У

ході побудови такого вузла важливо створити правильну класифікацію атрибутів для того, щоб домогтися чистоти роботи вузла. Тобто, вузол має “розбивати” змінні таким чином, щоб до кожної підкатегорії відносилися тільки ті дані, які мають її характеристики, тобто щоб кожна одиниця даних спрямовувала у коректну підкатегорію.

“Випадковий ліс” же - це комплекс дерев прийняття рішень. Вхідні дані, у ході свого аналізу, проходять крізь низку дерев. Зважаючи на те, що даний метод машинного навчання застосовується не тільки для завдань регресії, але і для завдань класифікації, вихідні результати також будуть різні: для регресії вихідні значення будуть представлені у вигляді середнього, а для класифікації шляхом голосування визначається вид кінцевого класу.

Сильною стороною даного методу є гарна продуктивність і що він може бути рівноцінним замінником нейронних мереж. До того ж, даний вид регресії досить простий для розуміння, тому навіть для початківців цей вид машинного навчання буде зрозумілим.

Проте є і недоліки. Даний метод має здатність до “перенавчання” - тобто коли кінцева модель прийняття рішень стає дуже складною та містить велику кількість непотрібних даних та недоцільних структурувань. І ще варто зазначити, що дерева прийняття рішень та метод “випадкового лісу” - є дуже часозатратним та потребує великої кількості оперативної пам’яті комп’ютерного пристрою, з якого відбувається моделювання.

Також, якщо говорити про регресію в цілому, варто зазначити, що вона застосовується не тільки в завданнях прогнозування та виокремлення якісних тригерів передбачення, але й пошуку аномалій. Завдяки впорядкованості своєї структури, даний метод є демонстративним показником девіантних даних, що також є цікавим полем вивчення для соціологів:

“Коли в нас є певний ряд, ми намагаємось його передбачити і знайти ідеї відхилення нашого передбачення і таким чином виявити аномалії і спрогнозувати якого типу ця аномалія - це у нас іде якийсь тренд чи це одиничний скачок який потім прибереться” (Ганна, 5)

Отже, завдання регресійного аналізу у соціологічному дослідженні може бути виконано за допомогою машинного навчання і для цього існують декілька різних моделей зі своїми специфічними особливостями. Загалом регресійна модель є модифікованою версією класифікаційної, тому доцільно їх вивчати саме у наведеній послідовності.

2.4. Використання алгоритмів машинного навчання для завдань контент аналізу

У сучасному соціологічному аналізі документів необхідним є притягання до дослідження великої кількості людських ресурсів і у той же час зменшення до мінімуму впливу людського фактору на результати дослідження. У випадку з класичним контент-аналізом, такі дослідження застосовуються як локальні, адже мають на меті вивчення окремих видів документів (зображення, текст, аудіо тощо). До того ж, з огляду на конвенційний спосіб підбору досліджуваних документів, результати дослідження можуть бути не репрезентативними. Таким чином можна дійти до висновку, що суб'єктивний характер побудови вибірки та вибір способу висвітлення досліджуваного явища може суттєво вплинути на відображення соціальної дійсності та викривити результати дослідження.

Проте, цього можна уникнути, впроваджуючи методи машинного навчання у процес аналізу текстів, а також аби розширити поле та матеріал вивчення.

Ось як характеризує перспективи інтегрування машинного навчання у сучасний контент-аналіз в Україні респондентка Христина Снопик:

"Насправді роботи «неорене поле» і це все мабуть залежить від того, що чим більше спеціалістів, тим більше попиту і т.д. Звісно, що працювати багато над чим, всі країни, які хочуть підтримувати розвиток своєї мови, розвиток автоматичного опрацювання мови, вони будуть над цим працювати. Мені здається, що це тільки природнє власне розширюватись в цій справі." (Христина, 4)

Новітні технології машинного навчання, що наразі вдосконалюються та розгалужуються розробляючи новий функціонал поширюють практику

переводу різного роду контенту (друковані книги, документи, газети тощо) у цифровий формат. А це, в свою чергу, надає дослідникам нові можливості для контент аналізу (Денисенко, Чеботарьова, 2008).

Проте як класичний контент аналіз, так і аналіз з використанням машинного навчання потребує особливого підтвердження валідності моделей та їх перевірку на коректність.

“Класичний КА доцільно застосувати до даних, які не є великими, не потребують суцільної автоматизації кодування. КА із використанням МН не вимагає перевірки надійності та відтворюваності кодування, але, натомість, вимагає особливої уваги до валідності створених моделей. Суто числові показники якості кодування не завжди адекватно показують придатність моделі для класифікації абстрактних сутностей, таких як гумор, наприклад. Це найголовніші, на мою думку, методологічні відмінності.”(Олег, 3)

Спираючись на так звані токени - заздалегідь сформовані одиниці синтаксичного виміру, є можливими розпізнавання великих масивів контенту, його класифікація за певними ознаками та впорядкування в залежності від смислового наповнення та кластеризації в залежності від розташування на системі координат.

Важливою складовою контент-аналізу за допомогою машинного навчання є етап обробки природної мови (NLP). Для кращого розуміння специфіки роботи цього етапу та співвідношення з машинним навчанням, до обговорення була запрошена спеціальна експертка, комп’ютерна лінгвістка Христина Снопик.

Для обробки текстових матеріалів зазвичай необхідно володіти (або створювати словник, завдяки якому є можливим інтерпретація тексту

алгоритмом. Він містить список слів, які можуть бути використані у тексті. У ході інтерпретації тексту слова переводяться у більш зрозумілу для алгоритму мову - бінарний код.

Подібні словники охоплюють множину використаних у досліджуваному тексті слів, що дає можливість їх розпізнати у текстових файлах.

“«Яку роль машинне навчання грає в NLP?» Останніми роками величезну. До того, для аналізу тексту, для перекладу, наприклад, для виправлення помилок, для класифікації тексту, для визначення теми тексту, найбільше використовувались якісь статистичні методи. Після 2007 після буму глибинного навчання, машинне навчання займало все більш і більш кращі позиції в опрацюванні природної мови, тому що NLP - це Natural Language Processing.” (Христина, 4)

Варто зазначити, що NLP - це окрема сфера зі своєю специфікою, що ближче до побудов правил та лінгвістики, ніж до соціології, проте без цього етапу неможливе здійснення контент-аналізу та багато інших, зазначених вище завдань. Проте, якщо говорити про специфіку саме українського контексту, є сенс говорити саме про специфічні риси української мови, які необхідно враховувати при роботі з вищезгаданими словниками. Як зазначала експертка, словників з англійською мовою створено вже досить багато, і ці бібліотеки постійно доповнюються, а таким чином можливе вирішення більш специфічних задач з залученням мінімальної кількості ресурсів. В українській мові ж склалася інша ситуація - можна побачити стрімке збільшення інтересу до машинного навчання в Україні з боку науковців лише починаючи з 2010 року, тобто матеріальна база українських вчених набагато менше та наразі тільки розвивається. Тому доречним є згадати основні особливості української мови як такої, на які треба зважати

при побудові правил обробки текстових файлів. І першим з них є неструктурованість частин мови у реченні, відсутність чітких правил побудови речення.

“По-перше, українська мова дуже флексивна мова, це означає, що в українській мові порядок слів не визначений, тобто підмет присудок, я можу сказати: «Я люблю каву» чи «Люблю я каву». Тому що в українській мові ми визначаємо відношення від слова до слова, що таке підмет, присудок, ми визначаємо відмінками «люблю» - 1 особа однини, ми визначаємо закінченнями, «каву» - «у» - закінчення відмінку «Кого? Що?» знахідного відмінку.” (Христина, 4)

Проте необхідно пам’ятати також про ще дві особливості української мови - це омонімія та синонімія.

Синонімія - це особливість мови, яка виражається у можливості вираження певної думки різними способами без зміни її смислового навантаження. Наприклад, слова “переміг” та “подолав” мають однакове значення і обидва можуть бути застосовані у реченні у тому ж значення.

У випадку омонімії є складність інтерпретування слова з огляду на його контекст, а це тягне за собою створення додаткових правил роботи з подібними частинами мови, адже значення слова варіюється згідно до контексту. Найпоширеніший з прикладів - слово “коса”, що може мати значення як виду зачіски для довгого волосся, мілкого місця у водоймі або знаряддя для сільськогосподарської роботи.

“А по-друге є так звана мовна омонімія, тобто можна сказати: «Замок». Що маємо на увазі? Який замок? Одним словом це складно визначати, якщо є однакове слово, тим паче слово у якомусь відмінку може означати 2 різних слова. Це додає складнощів. Тому з українською складніше

працювати." (Христина, 4)

І ще однією особливою рисою не української мови, а радше українського контексту використання машинного навчання є важкодоступність використання вже зібраного банку даних та неможливість їх отримання. Наприклад, недоступність використання даних Інституту української мови, що при Академії наук. До того ж, про це казав і інший респондент (Тимофій, 1), що наразі в Україні наявні масиви даних які до того ж постійно генеруються та доповнюються, проте вони зберігаються у такому форматі, що ними дуже складно оперувати (наприклад, скан роздрукованого документа).

"Я б назвала ще одну проблему, що є Інститут української мови при Академії наук. Кажуть, що вони мають хороший корпус даних, проанотований корпус української мови, там де частини мови проанотовані, зв'язки між словами, але це закритий ресурс. Принаймні рік назад ресурс був закритий. Чи бачив хтось цей ресурс, не знаю. Така ще бюрократична річ." (Христина, 4)

Тепер варто розглянути основні риси побудови подібних правил та їх характеристики. Так, як NLP є одним з етапів дослідження з використанням машинного навчання та є характерною рисою дослідження саме текстового матеріалу, вони мають свої особливості. Першим етапом є розпізнавання алфавітного ряду на основі певного набору характеристик кожної букви окремо. Головна особливість у тому, що алгоритм спочатку навчається розпізнавати кожну букву, потім ідентифікувати характеристики мови і визначає, чи відповідають ці дані тим, які він навчений аналізувати:

"Спочатку дивимось не так. До нас приходить величезна стрічка, рядок. Ми не знаємо, чи це англійська українська, чи ще щось, спочатку ми

розпізнаємо текст. Є бібліотеки, які вже натреновані, теж мають модельки, вони розпізнають тексти. Відповідно кажуть: «Українська-90%, російська-7% і т. д.». І ми визначаємо, що 90% це українська, значить це український текст і він нам підходить, якщо ми, наприклад, працюємо з український текстом." (Христина, 4)

Далі йде, власне, класифікація тематики або будь-якого іншого зазначеного параметру для інтерпретування стилістичної або тематичної направленості тексту. Схематично цей процес можна зобразити так: алгоритм класифікує всі наявні слова за частотністю та виявляє найбільш повторювані з них. Далі інтерпретує їх значення та згруповує в одну тематику, яка їх об'єднує. Так, наприклад, якщо у тексті найчастотнішими словами є "стать", "стамбульська", "віктімізація", то цей текст буде інтерпретований як той, що має тематику домашнього насилля.

А згодом, коли процес визначення тематики тексту завершений, можливе впровадження певних правил роботи з даними токенами. У даному прикладі пані Снопик описується правило виявлення та маркування орфографічно неправильного слова:

"Ми беремо найчастотніші слова і таким чином класифікуємо текст в якусь тему. Коли ми хочемо виправити помилку, то це трохи складніше. Беремо речення, вже не з токенами працюємо, а з реченням, наприклад, пише «будь-ласка» через дефіс (зараз в українській мові «будь ласка» без дефісу має писатись). По суті ми можемо такі правила написати: «якщо між «будь» і «ласка» дефіс – підкреслювати це слово». Таких правил можна написати багато, але їх довго писати."(Христина, 4)

Ще одним завданням, який можливо втілити за допомогою машинного навчання, є мережевий аналіз. Як вже було зазначено у попередніх

підрозділах, він є важливою складовою вивчення динамічних реакцій певних категорій суспільства на певні явища та події у сучасному світі. Варто зазначити, що чотири з п’яти експертів зазначали його важливість в дослідженні громадської думки, соціальних явищ та груп. Наразі такий вид аналізу є досить комерційним, адже завдяки ньому великі компанії, які можуть собі дозволити дослідження з використанням машинного навчання, вивчають свій репутаційний вплив на потенційну цільову аудиторію, знаходять нові характерні риси своєї аудиторії, тощо.

“Контент аналіз, наприклад, в соцмережах, дуже багато різних компаній великих, філій і т.д., вони заходять в соцмережі і по тому де вживається тег їх, вони аналізують, чи це позитивний відгук про їх техніку чи негативний. Це не тільки соцмережі, вони по всіх форумах «скреблять» дані і т.д.” *(Христина, 4)*

У ході розгляду інтегрування машинного навчання у контент-аналіз, перш за все, варто говорити про можливість застосування машинного навчання до розпізнавання та класифікації образів, що було продемонстровано вище. Проте для цього підходять декілька методів машинного навчання, і це надає дослідникам певну перевагу у вигляді варіативності у методі аналізу даних. За допомогою відповідей експертів було виокремлено декілька найбільш поширених та визнаних у колах дослідників методів:

LDA, або латентне розміщення Дирихле, або так зване тематичне моделювання. (Blei, Andrew, Jordan, 2003). Даний метод виконує схожий процес аналізу тексту, який був згаданий вище. У процесі його роботи згадані в тексті токени на вхідному шарі після процесу класифікації об’єднуються у смислові групи і завдяки спільній появі у одному й тому ж

документі визначають тематичну спрямованість тексту. До того ж, одне слово може належати до декількох смислових кластерів, при тому мати у кожному з них різне значення. У цьому випадку структурним елементом аналізу тексту виступають семантичні особливості токена та контекстуальні особливості його появи. Ключовим аспектом його роботи є те, що він працює завдяки статистичній моделі морфологічної побудови мови і дозволяє сформувати нову інтерпретацію тематики без впливу суб'єктивної або заангажованої інтерпретації дослідником. У межах цього методу можливий також ретроспективний аналіз соціальних взаємодій або явищ як, наприклад, дослідження сприйняття та критика політичних лідерів тощо. Важливо розуміти також, у якому випадку варто використовувати цей метод, а саме:

"LDA — зручно для експлораторного тематичного аналізу, коли не знаєш, з якого кінця підійти до формування категоріальної схеми." (Олег, 3)

SVM, або так званий метод опорних векторів. Ще один розповсюджений, навіть базовий метод класифікації за допомогою машинного навчання. Головна ідея цього методу заключається у створенні так званої гіперплощини, на якій розміщуються дані та групуються завдяки розділенню цих даних (проведенням лінії) на групи найбільш оптимальним способом. Назва методу походить від однієї з його функціональних характеристик - знаходження найближчої до лінії розмежування точки, що має назву опроного вектору. Після цього, алгоритм прораховує відстань від опорного вектору та роздільної площини. Ця відстань зветься проміжком, і основна функція алгоритму - це зробити даний проміжок якомога більше.

"SVM — підходить, якщо «зразки» мають одноманітну лексику і потрібна чітка класифікація за набором категорій." (Олег, 3)

XGBoost (eXtreme Gradient Boosting), або екстремальне градієнтне підсилення - це алгоритм машинного навчання, що базується на принципі

бустінгу за допомогою архітектури градієнтного спуску. XGBoost - це быблыотека для різних мов програмування, яку можна завантажити на комп'ютер та отримати до нього доступ через багато інтерфейсів (в т.ч і Python або R).

"XGBoost — ідеально для «розмитих» категорій за умови наявності великої бази класифікованих текстів (напр., позитивні / негативні коментарі до товарів у ел. магазинах). Чим більша база для навчання, тим краще. Чим більші обчислювальні потужності, тим краще теж. Розваги «по дорогому»." (Олег, 3)

Word2Vec та BERT . Word2Vec - це сукупність методів на основі штучних нейронних мереж. Дана модель визначає статистичний розподіл появи слів у тексті, шляхом застосування штучних нейронних мереж знижує їхню розмірність та розташовує у вигляді векторного графіку, що демонструє відношення слів у тексті.

BERT- це метод обробки природної мови, що заснований на використанні нейромереж для виконання завдань послідовності. Наприклад, завдяки цьому методу пошуковий запит у Google обробляється не тільки за прямим значенням, але й враховуючи контест слова, що дозволяє пошуковій системі видавати більш релевантні для користувача результати.

"Word2Vec і BERT — ідеально для класифікацій коротких текстів (напр., категорій товарів, видів діяльності, тощо)." (Олег, 3)

Отже, у даному підрозділі були зазначені перспективи використання методів машинного навчання у контент-аналіз в Україні. На основі відповідей респондентів були окреслені головні особливості та структуру даного методу, а також надані практичні рекомендації щодо його застосування.

2.5. Моделі застосування машинного навчання у контексті сучасної

методології в Україні

У ході дослідження експертами було висунуто низку рішень, які можуть допомогти соціологу з наявним рівнем знань та навичок почати вивчати нові, перспективні методи аналізу даних у вигляді машинного навчання. Проблема полягає в тому, що певна кількість дослідників в Україні зосереджена більше на теоретичній сфері соціології як дисципліни і вона розвивається значно швидше за її методологічну сторону. Саме тому у цьому дослідженні зібрані практичні рекомендації щодо застосування машинного навчання у соціології та розроблена чітка схема дій дослідника для проведення дослідження з використанням машинного навчання. Рекомендації зібрані таким чином, що навіть люди без жодного теоретичного та практичного досвіду матимуть змогу ними користуватися. На основі зазначених у попередніх розділах матеріалів, дані рекомендації є уніфікованими під будь-який досвід дослідника.

Разом з експертами також було розроблено певний специфічний алгоритм застосування подібних методів з точки зору завдань дослідження. Адже питання застосування даних методів щільно пов’язане з методологічною та теоретичною рамкою самого дослідження, тому є доцільним окреслити усі його стадії від вибору теми до інтерпретації результатів.

Першим етапом, як і в класичному академічному досліджені, є розробка схеми дослідження. Опис методологічної рамки, збір теоретичного матеріалу, формування мети, об’єкту та предмету дослідження, розробка дослідницьких завдань тощо. На цьому ж етапі є доцільним розробка методологічної схеми:

“Розробка концептуальної схеми дослідження: категорій, одиниць аналізу, одиниць рахунку (якщо категорії аналізу дискретні).” (Олег, 3)

Наступним етапом є розрахунок вибірки. У випадку з машинним навчанням потрібно зважати на те, що обсяг вибірки має бути набагато більший ніж в класичному досліджені.

Надалі дослідник має приступати до розробки методів обробки цих даних. Це напряму залежить від вже сформованої схеми дослідження, адже метод обробки має дозволяти досягти кінцеву мету дослідження та забезпечити репрезентативність результатів в залежності від сформованої вибірки.

Також на цьому етапі необхідно провести поглиблений аналіз можливих джерел даних, сформувати схему структуризації цих даних (якщо потрібно) та описати ключові моменти подальшого збору інформації з зазначеного джерела. Важливо пам’ятати, що необхідність збору інформації з кожного джерела має бути достатньо обґрунтованим та аргументованим, варто також подбати про захист персональних даних користувачів (якщо використовуються дані конкретної групи індивідів), а також ознайомитися з основними політиками обробки даних, що діють на території України та на територіях, які є потенційно цікавими для збору інформації.

“У випадку «великих» даних: підготовка інфраструктури для збору та обробки даних.” (Олег, 3)

Після цього йде етап застосування моделі збору інформації та їх попередньої обробки.

Далі йде об’ємний етап аналізу даних, що включає у себе: розробку та навчання моделей машинного навчання, їх застосування, збір проміжних результатів статистичної обробки задля коригування моделі та “довчання” тощо. (Олег, 3)

Останнім етапом йде інтерпретація отриманих результатів, що замикається підвердженням або спростуванням певних гіпотез, формування висновків дослідження та пошук вирішення досліджуваної проблематики.

Проте якщо говорити про моніторингові дослідження, то процес дослідження включає у себе ще один додатковий етап:

“У випадку моніторингових досліджень: налаштування конвеєру даних (data pipelines), еволюції моделей / нейронних мереж.” (Олег, 3)

Звичайно, без додаткового теоретичного базису для дослідника буде непростим завданням вихід за межі звичної методологічної парадигми, тому у ході інтерв’ю кожен з експертів зголосився надати список рекомендованої літератури за темою, у якій виступав у ролі експерта (Додаток Г). Даний список літератури згрупований за тематиками та може виступати у ролі довіреного методологічного базису для подальшої роботи та вивчення.

Також з огляду на такі буденні обставини, як нерозуміння рівня необхідного забезпечення дослідника для використання машинного навчання, у додатку Д (Додаток Д) э рекомендації щодо технічного та програмного забезпечення для дослідників, які мають намір почати роботу з машинним навчанням.

Отже, у ході дослідження було сформовано модель інтегрування методів машинного навчання у класичне соціологічне дослідження. Дана модель може використовуватися як оглядовий матеріал та введення у тему машинного навчання.

ВИСНОВКИ

Сучасна соціологічна методологія в Україні як в академічному, так і в практичному полі діяльності, рухається у бік технічної адаптивності та міждисциплінарного використання всіх наявних засобів збору та аналізу інформації, що може надати сучасна науково-дослідницька спільнота. Соціальний простір заповнюється веб-простором та навпаки - соціальні взаємодії дедалі частіше пов’язані з активністю індивіда у мережі Інтернет. Кількість необроблених даних збільшується, як і потенціал їх використання для соціологічних досліджень, а зі збільшенням кількості даних та глобалізації питання вивчення соціально-культурних, демографічних ознак та споживацьких звичок, збільшується і перспектива екстраполювання результатів дослідження на світові масштаби.

Для розвитку соціологічної методології саме в Україні це має велике значення, адже дозволить сформувати певне уявлення про поточний стан розвитку та взаємодії в українському суспільстві, а також динамічно відслідковувати реакції суспільства на події в країні. Використання машинного навчання у вирішенні соціологічних задач може дати провести більш поглиблений аналіз українського соціального простору.

Проте використання методів машинного навчання у соціологічній методології передбачає кардинальну зміну вектору розвитку сучасних досліджень та нову, специфічну підготовку молодих фахівців у академічному середовищі. Саме тому у ході дослідження був систематизований категоріально-понятійний апарат науки “data science” та висвітлені основні моменти її складових для загального введення у дану тему.

Також задля розкриття теми доцільності інтеграції методів машинного навчання у сучасну соціологічну методологію були окреслені основні підходи до міждисциплінарного спрямування у сучасній науці, а також окреслена основна специфіка роботи з штучним інтелектом. Ця специфіка полягає у: зміні соціального простору та соціальної взаємодії, що виникла у результаті активного залучення індивідів у взаємодію з цифровим простором, проблема швидкості обробки даних, адже при сучасних темпах їх збирання вони можуть вважатися застарілими вже через досить короткий проміжок часу. Також має місце заангажованість соціологічної науки як такої, важкість доступу до певних видів даних та недостатній рівень технічної підготовки дослідників.

Також було висунуте припущення, що має місце певна методологічна специфіка побудови теоретичної обґрунтованості дослідження, що полягає у різному підході до формулювання теми дослідження. Тобто, класичне соціологічне дослідження виходить більше з засад висування гіпотез та їх перевірки, а спеціалісти з машинного навчання виходять скоріш одразу з наявних даних. Проте це припущення не було підтверджене експертами та був зроблений висновок, що має місце гнучкий підхід до обґрунтування вибору теми дослідження як у спеціалістів з машинного навчання так і в соціологів, тобто ці підходи в обох сферах можуть комбінуватися.

У ході дослідження було проведено ряд експертних інтерв’ю як з соціологами, що мають практичний досвід у використанні методів машинного навчання у своїх дослідженнях, так і з розробниками моделей машинного навчання, які надали більше рекомендацій щодо технічної складової використання даних методів. Базуючись на відповідях експертів було сформовано перелік основних задач соціологічного аналізу даних, у яких буде доцільно інтегрувати методи машинного навчання. Цими

завданнями стали: Класифікація та кластеризація, регресійний аналіз та контент-аналіз. Варто зауважити, що даний перелік є узагальненим та виступає оглядовим матеріалом для знайомства з темою та швидкій навігації в ній.

Результатом цього дослідження виступає розроблена покрокова інструкція з загальними рекомендаціями по проведенню соціологічного дослідження з використанням машинного навчання зважаючи на український контекст, сформований перелік корисних матеріалів для поглиблення у кожну з описаних у роботі тематик та наданий перелік можливих джерел великих даних.

Отже, у ході дослідження було розглянуто питання інтеграції машинного навчання у сучасну соціологічну методологію в Україні. Був сформований допоміжний оглядовий матеріал для соціологічних дослідників для першого ознайомлення з перспективами інтегрування методів машинного навчання для вирішення їхніх дослідницьких завдань та обґрунтована доцільність даної інтеграції.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

Буданов, В.Г. (2007). 5.3. О курсах “Синергетика для гуманитариев” *Методология синергетики в постнеклассической науке и в образовании. (*сс. 174-210). Москва: Прогресс-Традиция.

Вайгенд. А. (2017). *Big Data. Вся технология в одной книге* (пер.з англ. Богданов С.) (ст. 27). Видавництво “Эксмо”.

Волков В. В., Скугаревський Д.А., Титаєв К.Д. (2016.). Проблемы и перспективы исследований на основе Big Data (на примере социологии права). *Социологические исследования №1, 2016*. ст. 48-58.

Волович, В. І. (2010). Методологічні проблеми соціологічного дослідження. *ВІСНИК Київського національного університету імені Тараса Шевченка* (1-2/2010).

Волович, В. І. (1974). *Надежность информации в социологическом исследовании.* Київ: Наукова думка.

Горбань, А. (1990). *Обучение нейронных сетей*. Москва:СССР-США СП "Параграф"

Дюркгейм, Е. (1897). У Чеховський Г.М., Грішина Л.П., (ред). *Самоубийство. Социологический Этюд.* (с. 20). Москва: Мысль.

Касавин, И. (2004). Философия познания и идея междисциплинарности. *Эпистемология и философия науки. 2(2)* (сс. 5–14.).

Корсунська, М.В., Семенова, Л. А. (2010). *Контент-анализ СМИ: проблемы и опыт применения.* (сс. 5-19) Москва: Інститут соціології РАН.

Паніотто, В. (1986). *Качество социологической информации.* Київ:

Наукова думка.

Поляков, Г. (1965). О принципах нейронной организации мозга. *Издательство Московского университета.*

Ядов, В. 5. ВЫДВИЖЕНИЕ РАБОЧИХ ГИПОТЕЗ. *Социологическое исследование: методология программа методы* (с. 40).

All children to receive whole genome sequencing at birth, under ambitions laid out by Matt Hancock. (5 November 2019). Retrieved from https://www.telegraph.co.uk/news/2019/11/05/children-receive-whole-genome-sequencing-birth-ambitions-laid/amp/

Alreck P.L., Settle R.B. (1985). *The Survey Research Handbook. Homewood, IL: Dow Jones-Irwin.* Homewood.

Annalyn Ng, Kenneth Soo. (2017). 6. Regression Analysis. В *Data science for the Layman.*

Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol. (2016, March 12). Retrieved from https://www.bbc.com/news/technology-35785875

Barlow, M. (2013). Real Time Big data Analytics: Emerging Architecture. *O`Riley Media*.

Cai T., Zhou Y. (2016). *What should sociologists know about big data?* University of Macau.

Carmines E.G., Zeller R.A. (1979). *Reliability and Validity Assessment.* Beverly Hills, California: SAGE.

Cielen D., Arno D. B., Meysman, Ali M., (2016). 3. Machine learning. В *Introducing Data Science* (p. 58).

EMC Education Services. (2015). *Data Science & Big Data Analytics.*

Canada.

Hastie T., Tibshirani R., Friedman J. (2009). *The Elements of Statistical Learning.*

IDC. (2017). IDC`s Data Age 2025 study, sponsored by Seagate. *IDC`s Data Age 2025 study, sponsored by Seagate.* Retrieved from https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf

James G., Witten D., Hastie T., Tibshirani R. (2017). *An Introduction to Statistical Learning.* Springer. DOI 10.1007/978-1-4614-7138-7

Knuth, D. (1974). Computer Programming as an Art. Retrieved from http://www.cs.bilkent.edu.tr/~canf/knuth1974.pdf

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 (27 April 2016) . *Official Journal of the European Union.* Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Results of the UNSD/UNECE Survey on organizational context and individual. (February 2015 г.). Retrieved from https://unstats.un.org/unsd/statcom/doc15/BG-BigData.pdf

Shai Shalev-Shwartz, Shai Ben-David. (2014). Understanding Machine Learning: From Theory to Algorithms *Cambridge University Press.*

Tilburg University (2019). History, Evolution and Future of Big Data. *British Journal of Management*, pp. 229–251. DOI 10.1111/1467-8551.12340

Wallerstein. I. (1987). World-Systems Analysis Social Theory Today. In I.Wallerstein, *World-Systems Analysis* A. Giddens., J.H.Turner, (Trans), pp. 309 - 324.). Cambridge: Polity Press.

25 Highest Paying Jobs in America. (2019). Retrieved from https://www.glassdoor.com/List/Highest-Paying-Jobs-LST_KQ0,19.htm

Gelman, A., John B., Carlin, Hal S. Stern, D.B. Rubin. (2003). *Bayesian Data Analysis* (Second edition). CRC Press.

Додаток А

Гайд для проведення інтерв'ю з експертом з соціологічних методів

ПІБ експерта _____

Посада _____

Метод _____

Дата проведення інтерв'ю _____

Блок 1. (Особистий досвід респондента)

- Скажіть, яку посаду Ви займаєте на даний час?
- У чому Ваш науковий інтерес?
- Що саме на стику соціології та машинного навчання Ви вивчаєте?
- Який Ваш досвід використання даних методів?
- Чи не могли б Ви коротко окреслити принцип його роботи?

Блок 2. (Перспективи інтегрування в соціологію)

- На вашу думку, у яких сферах буде доцільно використовувати методи машинного навчання?
- Скажіть, чи бачите Ви перспективи інтегрування методів машинного навчання у сучасну соціологічну методологію?
- *Якщо так:* які?
- *Якщо ні:* чому?
- Чи є перспектива інтегрування методів машинного навчання у сучасну соціологічну методологію?
- *Якщо так:* Окресліть, будь ласка, які методи машинного навчання, на

Вашу думку, можна використовувати в соціологічній практиці? Чому?

- *Якщо ні:* чому? *(після відповіді завершення розмови)*
- Які можуть бути проблеми у використанні цих методів в українських реаліях?
- На Вашу думку, де досліднику взяти чистий масив великих даних?
- На Вашу думку, чи доцільно досліднику знати певні мови програмування для самостійного створення алгоритмів?
- *Якщо так:* Які?
- *Якщо ні:* Чому?
- Можливо є ще щось, про що б Ви хотіли розповісти стосовно цієї теми?
- Чи не могли б Ви після цього інтерв'ю поділитися рекомендованою літературою для вивчення дослідниками, які мають хочуть почати аналізувати дані за допомогою машинного навчання?

Додаток Б

Гайд для проведення інтерв'ю з експертом по окремому методу

ПІБ експерта _____

Посада _____

Метод _____

Дата проведення інтерв'ю _____

Блок 1 (Теоретична інформація на основі власного досвіду)

- Скажіть, яку посаду Ви займаєте на даний час?
- У чому заключається Ваша робота?
- З яким видом машинного навчання Ви працюєте?
- Як саме Ви застосовуєте машинне навчання у своїй роботі?
- Чи не могли б Ви коротко окреслити принцип його роботи?
- Які специфічні особливості цього методу?
- Які проблеми з ним виникають?
- Яке їх вирішення?

Блок 2. (Перспективи інтегрування в соціологію)

- На вашу думку, у яких сферах буде доцільно використовувати методи машинного навчання?
- Чи бачите Ви перспективи інтегрування методів машинного навчання у сучасну соціологічну методологію?
- *Якщо так:* які?
- *Якщо ні:* чому?
- Чи є перспектива інтегрування методу (*метод, у якому експерт*

компетентний) у сучасну соціологічну методологію?

- *Якщо так:* Окресліть, будь ласка, як Ви бачите механізм його роботи.
- *Якщо ні:* чому? *(після відповіді завершення розмови)*
- Які можуть бути проблеми у використанні цього методу в українських реаліях?
- Можливо є ще щось, про що б Ви хотіли розповісти стосовно цієї теми?
- Чи не могли б Ви після цього інтерв'ю поділитися рекомендованою літературою для вивчення дослідниками, які мають хочуть почати аналізувати дані за допомогою *(назва методу, у якому експерт компетентний)*

Додаток В

Транскрипт інтерв'ю з Олегом Івановим

Дата проведення: 05-07 травня

Тривалість інтерв'ю: на вимогу респондента було проведено у письмовому форматі

Мова респондента: українська

І.: Скажіть, яку посаду Ви займаєте? У чому заключається Ваша робота як дослідника?

Р.: На сьогодні я не займаю ніяку «посаду», оскільки не працюю у штаті якоїсь організації. Виконую роботу на замовлення за прямими контрактами із фіксованою оплатою (рідше) або на умовах погодинної оплати. У загальному мою роботу можна окреслити такими напрямками: текст майнінг, програмування систем збору та обробки текстових даних (text mining, data mining & management, data extraction / ETL), кількісний аналіз текстових даних, у т.ч. із застосуванням машинного навчання, програмування роботів для збору, трансформації та поширення контенту на сайтах соціальних мереж (Facebook, Telegram, Instagram, YouTube, тощо).

Не все із цього є дослідницькою роботою і є ближчим до Data Engineer, ніж до Data Scientist.

Як приклади суто дослідницьких проектів можна навести аналіз контенту анімаційних фільмів за метаданими у професійних кінематографічних базах з метою виявлення популярних типів персонажів, тематик, тощо, на замовлення одного продюсера; розробка системи виявлення у соціальних мережах коментарів із мовою ненависті; аналіз різних форм насильства в українській публічній політиці із розробкою моделі виявлення текстів про

насильство та подальшому її застосуванні до бази із понад 1 млн. публікацій в укр. Інтернет ЗМІ за період 2000-2019 рр.

І.: Чи застосовуєте Ви машинне навчання у своїй роботі? У яких випадках?

Р.: Найчастіше коли необхідно автоматично класифікувати великий обсяг текстових даних за категоріями, які не піддаються опису певним набором ключових слів. Наприклад: позитивні/негативні відносно певного об'єкта висловлювання, мова ненависті, «насильницькі», гумористичні тексти. Також ML є безальтернативним у випадку необхідності автоматичної класифікації зображень, аудіо- та відео- даних, але цим я займався лише в особистих цілях, замовлень таких не було.

І.: Чим контент-аналіз з використанням методів машинного навчання відрізняється від традиційного контент-аналізу?

Р.: Краще писати «класичний контент-аналіз», а не «традиційний контент-аналіз», щоб не плутати їх із «традиційним аналізом тексту», який є якісницькою методикою «аналітичного прочитання тексту» (це, наприклад, мають робити всі студенти, готуючись до семінарів). Засади класичного контент-аналізу були розроблені у середині XX століття Бернаром Берельсоном та Гарольдом Ласуеллом та з тих часів не змінилися.

Класичний КА доцільно застосувати до даних, які не є великими, не потребують суцільної автоматизації кодування. КА із використанням МН не вимагає перевірки надійності та відтворюваності кодування, але, натомість, вимагає особливої уваги до валідності створених моделей. Суто числові показники якості кодування не завжди адекватно показують придатність моделі для класифікації абстрактних сутностей, таких як гумор, наприклад. Це найголовніші, на мою думку, методологічні відмінності.

Суто технічних та процедурних відмінностей більше. МН вимагає навичок програмування не лише для створення моделі, а й для підготовчого етапу ETL, розуміння математичних принципів побудови тієї чи іншої моделі. Нерідко дані є «великими», що вимагає знань та вмінь для побудови інфраструктури збору та обробки даних, паралельних обчислень, тощо. У великих проектах необхідними є або закупівля дорогого комп'ютерного обладнання або аренда обчислювальних потужностей «у хмарі» (наприклад інстансів EC2 AWS).

Відмінностей досить багато, всі їх окреслити важко, треба обрати якийсь вимір відмінностей: методологічний, технічний, організаційний, тощо.

Також є і спільні риси, але питання цього не стосується, я так розумію. Про методологічні відмінності різних «контент-аналізів» рекомендую мою статтю: http://ekmair.ukma.edu.ua/handle/123456789/7244

І.: Які методи машинного навчання доцільно використовувати у контент-аналізі?

Р.: Всі, які підходять для відповідної задачі дослідження. Не можна сказати, що якісь із методів чи породжуючих моделей МН не підійдуть для текстів у широкому розумінні (тобто не лише «наборів букв», а й аудіо-, відео- даних, тощо).

Однак, на сьогодні є популярні, «трендові» методи та моделі: LDA, SVM, XGBoost, Word2Vec, BERT. Див. статтю: https://link.springer.com/article/10.1186/s13673-019-0205-6 Певний час (переважно у 2010х) у політологічному середовищі популярною була модель WordScores, яку за певними ознаками можна вважати моделлю керованого навчання (supervised learning), хоча насправді це радше окремий випадок аналізу відповідностей (correspondence analysis). Див. статтю: http://faculty.washington.edu/jwilker/559/Lowe.pdf

І.: Чи не могли б Ви коротко окреслити принцип їх роботи?

Р.: Взагалі, не міг би. Це досить складні методи, коротке визначення яких породжуватиме хибні інтерпретації, а фахівці розкритикують такі визначення (я не вважаю себе фахівцем із МН, я застосовую його лише як інструмент в роботі за відпрацьованими алгоритмами).

Але «щоб було»:

LDA — Latent Dirichlet allocation, латентне розміщеня Діріхле. У кожному тексті є деякий (число визначене апріорно чи за певним статистичним критерієм) набір тем, який можна визначити за спільним зустрічанням слів.

SVM — support vector machines, метод опорних векторів. Метод керованого навчання, «на вході» якого є одиниці аналізу (тексти), класифіковані за однією чи більше категорією, на виході — векторне представлення одиниць у просторі.

XGBoost — eXtreme Gradient Boosting, екстремальне градієнтне підсилювання. Бібліотека для різних мов програмування, яка дозволяє організовувати розподілене обчислення моделей градієнтного підсилання на великих даних. Вихідною точкою є набір (градієнт) дерев рішень, які комбінуються в одну модель.

Word2Vec — сукупність моделей, побудованих на основі нейронних мереж, що представляють слова у корпусах природної мови у векторному вигляді. Застосовується, напр. при передбаченні слів у наборі з телефону. Тобто це динамічно змінювані ймовірності послідовностей слів. Рекомендую читати https://habr.com/ru/post/446530/

BERT — контекстуалізований варіант Word2Vec, який дозволяє будувати вектори не лише слів, а й фраз. https://habr.com/ru/post/436878/

І.: Які специфічні особливості цих методів? ***(на основі власного досвіду***

роботи або вивчення цих методів)

Р.: Дуже коротко і це не методологічний аналіз, а особливості практики застосування.

LDA — зручно для експлораторного тематичного аналізу, коли не знаєш, з якого кінця підійти до формування категоріальної схеми.

SVM — підходить, якщо «зразки» мають одноманітну лексику і потрібна чітка класифікація за набором категорій.

XGBoost — ідеально для «розмитих» категорій за умови наявності великої бази класифікованих текстів (напр., позитивні / негативні коментарі до товарів у ел. магазинах). Чим більша база для навчання, тим краще. Чим більші обчислювальні потужності, тим краще теж. Розваги «по дорогому».

Word2Vec і BERT — ідеально для класифікацій коротких текстів (напр., категорій товарів, видів діяльності, тощо).

І.: Які можуть бути проблеми у використанні цих методів в українських реаліях?

Р.: Ніяких методологічних проблем немає. Українські тексти є принципово такими ж послідовностями символів, що і англійські.

Щодо організаційних та ін. позаметодологічних обмежень — див. блок 2.

І.: Чи бачите Ви перспективи подальшого інтегрування методів машинного навчання у сучасну соціологічну методологію?

Р.: Тут скоріше доцільно говорити про «позитивні» і «негативні» перспективи. У мене досить радикальна позиція: соціологи або почнуть використовувати весь спектр сучасних кількісних методів, або соціологія як наука помре і перетвориться на «науковий комунізм» — псевдонауку, яка обслуговує ідеологічно ангажовані групи інтересів.

Тому далі я окреслю деякі важливі, на мою думку, фактори, які сприятимуть та гальмуватимуть таку інтеграцію. Наразі, нмд, більше

останніх.

І.: Які перспективи інтегрування методів машинного навчання у сучасну соціологічну методологію?

Р.: В усьому світі опитувальні та інші контактні методи соціологічних досліджень переживають кризу. Неухильно падає частка відповідей (response rate). У західних країнах F2F майже «померли», в Україні цей процес йде у тому ж напрямку і він неухильний. Телефонні опитування та опитування через Інтернет, нмд, є лише тимчасовим рішенням. В умовах інформаційного перевантаження, яке дедалі зростає, йде боротьба за увагу людини. Як зачиняють двері перед інтерв'юерами, так не будуть брати слухавку з невідомих номерів, так не будуть відповідати на електронні анкети.

Водночас, зростає обсяг «неспровокованих» персональних даних: записів у соцмережах, історій перегляду сайтів («кукі»), історій пересування (GPS трекінг), записів з камер спостереження, тощо. Це все вже аналізується. Не дивлячись на введення різних законодавчих обмежень, на кшталт GDPR, на практиці вони повсюдно порушуються. Всі ці дані є «великими» і можуть ефективно бути проаналізовані лише з допомогою сучасних кількісних методів, у т.ч. машинного навчання.

І.: У чому можуть бути складнощі?

Р.: Низький рівень технічної підготовки більшості соціологів. На сьогодні володіння, принаймні, однією мовою програмування є такою ж вимогою до професійної придатності на глобальному ринку праці, як і володіння англійською мовою. Застосування машинного навчання без написання коду можливе лише у готових програмних пакетах, які швидко застарівають та не забезпечують достатній рівень гнучкості для всіх дослідницьких ситуацій.

Недостатній рівень математичної підготовки. Сучасні методи МН є

досить складними з математичної точки зору і усвідомлене їх застосування вимагає відповідних знань.

Все ще сировинний характер української економіки. Інтелектуальний продукт не має такого рівня попиту на українському внутрішньому ринку, як на західних. Наш славетний IT-сектор — це на 90% аутсорс і аутстаф для західних компаній. На жаргоні політтехнологів, піарників і т.д. «зробити соціологію» значить провести опитування. Ні з якими іншими методами (тим більше неконтактними) соціологічні дослідження у широкому вжитку не асоціюються.

Моніторинги соцмереж, моделювання комунікативного середовища та ін. більш сучасні дослідження замовляють несоціологічним компаніям, які з досвіду, складаються із піарників, політтехнологів та IT-фахівців. У цих компаніях технофетишизм не підкріплений академічною базою у галузі соціальних наук. Однак на хвилі «хайпу» це досить добре продається, тому такі компанії не відчувають наразі потреби зближуватися із соціологічною наукою. Наприклад, 2 роки тому компанія LOOQME замовила мені розробку методики перевірки надійності їх даних кодування тональності, а також низки інших метрик. Замовлення було виконане і оплачене, але так і не імплементоване у «виробничий» процес.

Фахівці з МН, та й навіть з менш кваліфікованого Data extraction / ETL, є надзвичайно затребуваними в IT секторі. Соціологічні компанії не можуть конкурувати з ним за рівнем оплати праці. Та ж причина обумовлює недостатню кількість таких фахівців для викладання соціологам в університетах.

Сучасна соціологічна наука (у світі загалом) є ідеологічно заангажованою. Більшість наукових грантів мають «ціннісне» навантаження. Методологічні розробки з кількісних методів не вважаються вартими фінансової підтримки, бо їх «профінансує індустрія». Однак, у випадку МН

на практиці це не соціологічна «індустрія», а IT, куди і йдуть соціологи з відповідними знаннями і вміннями.

Є світовий тренд до зменшення можливостей «легального» збору даних для КА: GDPR, TOSи веб-сервісів, які забороняють автоматизований збір даних, урізання можливостей відкритих API. Всі це повсюдно порушують, але з іншого боку компанії, які накопичують ці дані (такі як Facebook та Google) всіляко намагаються технічно перешкодити їх отриманню дослідниками. Це «гонка озброєнь», яка в окремих випадках може вийти і в легальну площину з мільйонними позовами.

І.: Окресліть, будь ласка, як Ви бачите схему проведення контент-аналізу з використанням машинного навчання? ***(які кроки для проведення аналізу необхідно зробити досліднику, основні блоки дій).***

Р.: Якщо мова йде про академічне дослідження, то у цілому процедура не відрізняється від класичної, за винятком окремих технічних етапів.

Розробка концептуальної схеми дослідження: категорій, одиниць аналізу, одиниць рахунку (якщо категорії аналізу дискретні).

Розрахунок вибірки.

Написання програм для збору та попередньої обробки (ETL) даних.

У випадку «великих» даних: підготовка інфраструктури для збору та обробки даних.

Збір та попередня обробка даних.

Аналіз (розробка / навчання моделей, застосування моделей, проміжні статистичні аналізи, тощо).

Інтерпретація результатів.

У випадку моніторингових досліджень: налаштування конвеєру даних (data pipelines), еволюції моделей / нейронних мереж.

І.: Де, на Вашу думку, можна взяти достатню кількість даних (контенту) для дослідження з використанням машинного навчання в Україні?

Р.: В Інтернеті. Сайти соціальних мереж, публікації веб-ЗМІ, тощо.

І.: Яке для цього необхідне програмне забезпечення / навички роботи з ним?

Р.: Для контент-аналізу з використанням МН, починати треба не з вивчення власне МН та технічних засобів для нього, а з автоматизованого КА загалом. Наприклад, така, на перший погляд, тривіальна задача, як копіювання у базу даних заголовків із тисяч документів, може вимагати нетривіального знання мови регулярних виразів (regular expressions).

Починати варто із вивчення Python та базового набору бібліотек, таких як requests, re, beutufulsoup, pandas для ETL, scikit-learn для МН. Далі — R, бо сучасні методи кількісного аналізу у ньому імплементуються швидше, а деякі — є тільки у вигляді R бібліотек.

І.: Чи не могли б Ви після цього інтерв'ю поділитися рекомендованою літературою для вивчення дослідниками, які мають намір провести контент-аналіз за допомогою машинного навчання?

Р.: Див. мої дописи і коментарі у FB: https://www.facebook.com/groups/750366382143726/permalink/752985578548473/?comment_id=752995108547520&reply_comment_id=758362481344116

https://www.facebook.com/groups/750366382143726/permalink/786260398554324/

https://www.facebook.com/groups/750366382143726/permalink/763863410794023/?comment_id=765133000667064&reply_comment_id=765448540635510

Додаток Г

Транскрипт інтерв'ю з Христиною Снопик

Дата проведення: 06 травня

Тривалість інтерв'ю: 41 хв.

Мова респондента: українська

І.: Розкажіть, яку посаду займаєте, чим займаєтесь та який у Вас досвід роботи з машинним навчанням, з NLP і т. д.?

Р.: Я працювала с 2015 по 2019 рік, 4 роки в компанії «Грамарлі» на посаді комп'ютерної лінгвістики. Одразу скажу, що комп'ютерна лінгвістика включає в себе машинне навчання але це не є повністю машинне навчання, тому що в комп'ютерній лінгвістиці дуже багато статистичних методів використовується, та й просто на правилах. Після 2019 року працюю в маленькому стартапі «Байеслес», який займається виправленням упередженої мови (письмової мови) в англійській, зараз ми працюємо над українською «Вікіпедією». Найбільший я маю дослід, власне, з тим аби використовувати якісь стандартні бібліотеки комплінгвістичні, статистичні методи використовувати, з машинним навчанням я маю досвід, але небагато.

І.: Розкажіть, будь ласка, яку роль NLP грає в машинним навчанні та аналізу даних, яку планку вона займає?

Р.: Тут мабуть не так просто питання поставити. Яку роль NLP грає в машинним навчанні? Тут мабуть навпаки треба запитувати: «Яку роль

машинне навчання грає в NLP?» Останніми роками величезну. До того, для аналізу тексту, для перекладу, наприклад, для виправлення помилок, для класифікації тексту, для визначення теми тексту, найбільше використовувались якісь статистичні методи. Після 2007 після буму глибинного навчання, машинне навчання займало все більш і більш кращі позиції в опрацюванні природної мови, тому що NLP - це Natural Language Processing.

Я б не сказала, що воно витіснило повністю старші методи чи методи на правилах, але воно все більше займає кращі ланки і можна назвати переваги і недоліки обох методів:

- перевага машинного навчання в NLP, тому що не потрібно дуже багато лінгвістичної роботі, наприклад, аналіз тексту, якісь виписування лінгвістичних правил. Недолік в тому, що треба дуже багато текстів, дуже великий об’єм даних. Перевага старших методів, на правилах статистичних методів, там не потрібно багато даних, але потрібно багато лінгвістичної експертизи, людей з лінгвістичною освітою, і це досить довгий час забирає, власне аналіз текстів лінгвістичних і написання правил якихось.

Але я б не сказала, після 2018 або 2017 року, коли створили так звану модель «Берт» в машинному навчанні, машинне навчання займає все кращі позиції, але воно не витіснило зовсім ще якісь статистичні або комплінгвістичні методи і наврядчи витіснить насправді.

І ще один недолік є машинного навчання. Я казала про недолік машинного навчання, що там потрібні великі об’єми даних, і ще 1 недолік це те, що для тренування моделей необхідно багато обчислювальної техніки, це коштує дуже багато грошей, тому це теж ще один з недоліків

І.: Можете трошки розповісти для яких завдань NLP використовується і що воно дає на виході?

Р.: Добре. Самий стандартний приклад машинного навчання – всіма любимий google переклад – це використовує у техніки машинного навчання. Siri та Alexa чи ще щось – це теж NLP. Правда це вже не письмова, а розмовна мова, але це все ж те саме NLP.

Що ще? Виправлення помилок – Грамарлі теж саме, чи, наприклад, для багатьох мов є так зване Language Tool - це так саме NLP.

Класифікація текстів - це так саме NLP. Наприклад, коли заходити на якійсь новинний сайт, не всі новинні сайти використовують ручну систему тегування, просто в них є система яка розподіляє: спорт, політика, мода, краса і т.д.

Контент аналіз, наприклад, в соцмережах, дуже багато різних компаній великих, філій і т.д., вони заходять в соцмережі і по тому де вживається тег їх, вони аналізують, чи це позитивний відгук про їх техніку чи негативний. Це не тільки соцмережі, вони по всіх форумах «скреблять» дані і т.д.

Це найвідоміші випадки використання NLP.

І.: Можете трошки розповісти, як це реалізується? Як дослідник, що треба зробити, які блоки аби розпізнати текст як текст

Р.: Я працювала тільки з письмовим текстом, з власне усним мовленням я не працювала, тому я опишу pipeline, як проходить тексти від того початку**,** щоб його розпізнати і до кінця, щоб наприклад його перекласти, чи тег тег йому класифікувати. Спочатку дивимось не так. До нас приходить величезна

стрічка, рядок. Ми не знаємо, чи це англійська українська, чи ще щось, спочатку ми розпізнаємо текст. Є бібліотеки, які вже натреновані, теж мають модельки, вони розпізнають тексти. Відповідно кажуть: «Українська-90%, російська-7% і т. д.». І ми визначаємо, що 90% це українська, значить це український текст і він нам підходить, якщо ми, наприклад, працюємо з український текстом.

Прийшов величезний текст на 2 сторінки, ми розпізнали, що він на українській, далі нам треба розбити на якісь значущі частини. Це як зі школи учили – спочатку ми його розбиваємо на абзаци, потім на речення, потім можемо розбити на слова. В NLP найчастіше вживається термін «токен», а не «слово», тому що токен включає в себе слова, розділові знаки, цифри і т.д. Тобто, ми розбиваємо текст на абзаци, речення, токени і вже з такими значущими частинами ми можемо далі працювати. Якщо, наприклад, нам треба покласифікувати текст, то ми беремо і дивимось на найчастотніші слова, які вживаються в тексті. Якщо найчастотніші слова це «бігти», «перейти далі», то скоріш за все це спам, якщо це «мирний договір», «Франція», «Німеччина», то можливо це політика, і т. д.

Але я зверну увагу що все таки я говорю про методи комплінгвістичні, машинне навчання так не працює. Поясню трошки по іншому. Ми беремо найчастотніші слова і таким чином класифікуємо текст в якусь тему. Коли ми хочемо виправити помилку, то це трохи складніше. Беремо речення, вже не з токенами працюємо, а з реченням, наприклад, пише «будь-ласка» через дефіс (зараз в українській мові «будь ласка» без дефісу має писатись). По суті ми можемо такі правила написати: «якщо між «будь» і «ласка» дефіс – підкреслювати це слово». Таких правил можна написати багато, але їх довго писати.

Статистичний переклад колись використовували, фрази статистично підшуковували, до якоїсь української фрази, найчастіше вживану англійську фразу, і таким чином перекладали речення, але зараз статистичних перекладів нема, Google translator користується методом машинного навчання. Це такий pipeline, якщо ми візьмемо як працює дослідник з текстом для вирішення якихось проблем, якщо ми глянемо на те, як працювати з машинним навчанням з текстом, там дуже схоже, насправді, на початку. Ми беремо текст, визначаємо мову (українська, англійська і т.д.), розподіляємо його на абзаци, речення, токени, якщо нам потрібно, далі ми вже не пишемо ручних правил, ми просто запускаємо кожне речення в так звану функцію, яка визначає це речення ми пишемо правильно чи ні, ми можемо сказати, що в цій функції є так званий чорний ящик, це власне ці методи машинного навчання, які тренувались на великих обсягах даних і тепер знають, що «будь ласка» пишеться не через дефіс. Тобто можна ще не використовувати правила, використовувати так званий чорний ящик, цей чорний ящик можна розбирати і дивитись детальніше, як воно там все працює, але це здається вже не сюди.

І.: Скажіть, аналізу піддається усний і текстовий контент?

Р.: Так, усний і письмовий.

І.: Чи є якісь особливості з огляду на український текст, коли з огляду на контекст алгоритм розуміє, що там закінчення якесь не таке? Чи є такі особливості в українському тексті, українському контексті?

Р.: Є звісно. По-перше, українська мова дуже флексивна мова, це означає, що

в українській мові порядок слів не визначений, тобто підмет присудок, я можу сказати: «Я люблю каву» чи «Люблю я каву». Тому що в українській мові ми визначаємо відношення від слова до слова, що таке підмет, присудок, ми визначаємо відмінками «люблю» - 1 особа однини, ми визначаємо закінченнями, «каву» - «у» - закінчення відмінку «Кого? Що?» знахідного відмінку. В англійській мові відмінків нема, можна сказати є 2 відмінки, але відмінків все ж немає, тому сталий порядок слів, тому підмет має йти на початку: «I love coffee». Це спрощує дуже багато, тому що, на початку має йти підмет, присудок, таким чином правила зменшуються. Враховуючи що в українській мові закінчень дуже багато і порядок слів не сталий, набагато складніше опрацьовувати такі тексти, тому що складніше знайти зв’язки між цими словами, а по-друге є так звана мовна омонімія, тобто можна сказати: «Замок». Що маємо на увазі? Який замок? Одним словом це складно визначати, якщо є однакове слово, тим паче слово у якомусь відмінку може означати 2 різних слова. Це додає складнощів. Тому з українською складніше працювати. Якщо для англійської мови є дуже багато різних бібліотек і текстів, над якими вже працювали і корпусів, над якими вже працювали, і багато напрацьованих даних, то в українській мові такий даних мало. Це теж 1 проблема. Їх менше ніж в англійській, хоча останніми роками з’явились кілька інструментів, які допомагають з українською мовою працювати, зв’язки між підметом і присудком призначати, визначати частини мови і т.д. визначати. Є інструменти, вони не ідеальні, але вони хороші. Власне я зараз з ними працюю з українською «Вікіпедією»

І.: Є ще якісь проблеми з такою функціональної точки зору? Не загалом проблеми, які виникають, наприклад, з аналізом українського тексту, а відсутністю бібліотек, відсутність ще чогось, що заважає Україні

розвиватися.

Р.: Ви маєте на увазі, що заважає розвивати саме український NLP?

І.: Так

Р.: В принципі, мабуть, перше, що я назвала, даних мало, над ними працюють деякі команди людей, але все одно мало. Бібліотек кілька за останні кілька років з'явилися для розбиття тексту на абзаци, речення, слова, для позначення частин мови, для означення зв'язків між реченнями. Я б назвала ще 1 проблему, що є Інститут української мови при Академії наук. Кажуть, що вони мають хороший корпус даних, проанотований корпус української мови, там де частини мови проанотовані, зв'язки між словами, але це закритий ресурс. Принаймні рік назад ресурс був закритий. Чи бачив хтось цей ресурс, не знаю. Така ще бюрократична річ.

І.: Виникло таке питання. Аналіз тексту, на прикладі соцмереж, в інтернет просторі, який ми можемо скопіювати, розпізнати, а якщо це якісь матеріали, відскановані газети, там якийсь інший принцип роботи, тобто там же потрібно розпізнати спочатку по буквах, що це за буква і тому подібне, чи це якось відрізняється?

Р.: Ні воно не відрізняється, але залежно від того, що ви хочете з цим текстом. Єдине, що пам'ятаєте, що спочатку покроково розпізнаємо мову, чи це англійська, чи українська. Перед цим кроком, буде крок, як ві сказали, розпізнання букв, англійською воно називається optical character recognition, позначають OCR? Ця проблема вже вирішена, OCR дуже гарно розпізнає всі тексти. Є ще проблема, відповідно до правопису 2019 року «пів-Європи»

мало писатися через дефіс. І якщо, наприклад, в газеті було б написано «пів-» і «Європи» з іншого рядка, було складно розпізнати воно мало писатися через дефіс насправді, або має писатися разом і це просто перенесення. Але ці проблеми визначаються різними правилами. Є ще проблеми щодо українською, бо в англійській ці проблеми були вирішені. Проблема, як п’ятірку погано розпізнали, як літеру “s” англійську. Це не дрібниці, але так в принципі воно добре розпізнає.

І.: Якщо говоримо про приклад про «пів-яблука», наприклад, що ми можемо зробити потім з цією інформацією?

Р.: Теж саме, я згадувала, про компанію Philips, яка дивиться на тексти в соцмережах і на форумах. По-перше є багато напрацьованих різних семантичних словників, в цьому семантичному словнику кожне слово має позначку, яка означає, вона нейтральне, негативне, дуже негативне, позитивне, дуже позитивне. таким чином, проходячись по тексту з таким словником, ми можемо визначити, чи текст негативний загалом, чи ні. Є проблеми, як то часто «не» не враховуються. Але це теж вирішується. Машинне навчання також гарно позначає ці проблеми, так само добре позначає текст на позитивний чи негативний.

І.: Я правильно розумію, ми сприймаємо розпізнаваний текст, в якому ми орієнтуємося і можемо робити будь які задачі, наприклад, за допомогою того ж машинного навчання, розпізнавання особливостей діалекту, співставлення і тому подібне. Ми можемо використовувати весь функціонал машинного навчання і стільки, наскільки нам хватить грунтованності, доцільності?

Р.: Так

І.: Розкажіть будь ласка на якому етапі у нас в Україні вивчення, розвиток NLP?

Р.: Якщо говорити про комплінгвітику, то в Україні є кілька університетів, яку мають кафедру комплінгвістики. Найкращий університет з тих, що я знаю це КНУ ім. Шевченка. Він має кафедру комплінгвістики, там хорошу підготовку дають. Є ще Донецький університет в Вінниці, має кафедру прикладної лінгвістики, є в Львівському політехнічному, в Харкові. Тобто десь когось готують по комплінгвістиці, але в малій кількості і не дуже грунтовно. Тільки в КНУ, з того, що я чула, гарно готують. Це якщо говорити про освіту.

Грамарлі робить щорічні, кожного літа літню школу з комплінгвістики, тобто, якщо якісь спеціалісти з тих університетів хочуть щось ще довчити, вони можуть прийти на лютню школу. Це теж плюс. Я вже казала про Інститут української мови при Академії наук, він має якісь ресурси, опрацьовані.

Щодо індустрії, в нас є компанії, які займаються опрацюванням природного тексту, той самий Грамарлі, працює тільки для англійської. Є компанія яка займається семантичним аналізом мови соцмереж. Не дуже багато їх є, але все ж є.

І.: Взагалі, потрібно в Україні цьому вникати, як саме в українському контексті, не як в Грамарлі, там де більше аналізується англійський

текст, а не український. В Україні такому потрібно виникати, є якісь перспективи?

Р.: Насправді роботи «неорене поле» і це все мабуть залежить від того, що чим більше спеціалістів, тим більше попиту і т.д. Звісно, що працювати багато над чим, всі країни, які хочуть підтримувати розвиток своєї мови, розвиток автоматичного опрацювання мови, вони будуть над цим працювати. Мені здається, що це тільки природнє власне розширюватись в цій справі.

Мені б хотілось бачити розвиток.

Компанія YouScan, вони власне сканують тесті з соцмереж і дивляться, як їх аналізують. Я знаю, що компанія «Моршинська» надавала їм послуги, і вони дивились, як про цю компанію відгукуються.

«YouScan», Грамарлі – найвідоміші зараз.

Ви мене питали хто українську використовує, правильно? З українською на жаль небагато. Є новинний сайт «Ліга закон». Це новинно-аналітична компанія і вони опрацьовують українську і російську мови. Є ще компанія «Maklai», вона в Києві є, це як буклінг український, вони використовують NLP для української і російської. Це мабуть все, що я знаю про українську.

І.: Скажіть, де брати дані для наприклад машинним навчання, аналізу?

Р.: По-перше є такий корпус, який називається Державний золотий стандарт, де проанотовано українські частини мови і зв’язки між реченнями, це 1 корпус. Дані можна самим позбирати з, наприклад, новинних сайтів, з соцмереж, називається – «наскрейтити». Є ще «Вікіпедія» українська, стягнена з Інтернету, так саме можна стягувати дані з соцмереж і новинних

сайтів, так саме і «Вікіпедію» стягнули.

І.: Це якийсь парс, тобто просто пишеться спеціальний бот, який знаходить необхідні дані по заданим характеристикам?

Р.: Так, от саме так. Є такий не корпус, а більше інструмент «браунд юкей» називається, це проанотований словник, де проанотоване кожне слово, де і які частини мови можуть бувати і т.д., це волонтерський був проект.

І.: Він якось допомагає, яка його роль може бути у дослідженні?

Р.: Так, власне він допомагає редагувати слово, можуть бути проблеми, чи то знахідний або родовий відмінок, він допомагає власне омономію стягнути.

І.: Розкажіть, яка кількість людей мають працювати з якоюсь задачею? Це звичайно Ваш досвід, Ви працювали. Хто ці люди, яка їх спеціальність?

Р.: Задач просто є багато і вони різні. Наприклад, над цим волонтерським проектом, там ми вибирали тексти, корпуси, тексти для корпусу, потім проанотовували ці тексти. В волонтерський проект було залучено до 20 людей на постійній роботі з лінгвістичною освітою, тому що там власне були потрібні лінгвісти. Хтось робить якісь міні-проекти для української мови, як свою магістерську чи бакалаврську роботу. В КНУ так, наприклад, роблять. Це в принципі 1 людина. Дивлячись що треба, яку експертизу, якщо для лінгвістичного аналізу, то лінгвістичну експертизу, для того, щоб наскрейтити дані із соцмереж треба вміти гарно програмувати, програмісти

або програміст треба.

І.: Якщо ви використовуєте машинне навчання, хто цим займається, хто навчає, хто створює, чи є якісь готові продукти якими ви користуєтесь?

Р.: Ви питаєте, чи я зараз використовую машинне навчання?

І.: Взагалі, на Вашому досвіді, хто цим займається, якщо вам потрібно використати дослідження машинного навчання, ви забезпечуєте правила, хто займається навчанням алгоритму, створенням його, чи він вже створений, тобто якісь готові продукти є?

Р.: Цим займається, посада так і називається scientist, вона ще може називатись research scientist і різні варіанти. Це люди по сути, які вміють програмувати добри, мати хороші знання зі статистики. Лінгвістичні експертизи, не те що дуже потрібно, але вміння гарно програмувати і знання статистики.

І.: Яке потрібне програмне або технічне забезпечення для такої роботи? Ви казали, що необхідно багато техніки і програм для вашої роботи. Які це програми, яка це техніка, які стандарти?

Р.: Вже такі стандарти в машинному навчанні більше використовуються в мовах програмування Python, тобто треба вміти програмувати на Python. Треба вміти користуватися різними бібліотеками. Наприклад бібліотека “Стенфорд енелпі”. Це власне бібліотека, яка вже може протегувати і проаналізувати текст на зв’язки між словами, тобто вміти користуватись

бібліотеками. Є інші бібліотеки типу стандартної бібліотеки NLP Key**.**

Для того щоб добре працювати з машинним навчанням треба знати такі бібліотеки, як “scikit-learn” бібліотека Python, і “тензорфлоу” для того, щоб працювати з Python.

Що стосовно обчислювальної спроможності, то тут власне, сервер, вам дома його не потрібно мати. Ви просто купуєте місце на сервері, запускаєте віддалено свою програму, модуль, яка тренується на масиві даних. Вона може тренуватись від декількох годин до декількох днів. Вона віддалено тренується.

Це те, що треба з функціоналу.

І.: Ви використовуєте в своїй роботі Python, або так як ми більше про машинне навчання, ви використовуєте щось інше?

Р.: Не дуже зрозуміла питання. Так я використовую всі ці бібліотеки, пишу на Python, зараз працюю з бібліотекою «Stanza». Воно має готову модель, натреновану на українській мові. Машинним навчанням я зараз не займаюсь. Відповідно з бібліотеками, які стосуються машинного навчання, я не працюю з ними.

І.: Добре. Здається, це всі запитання. Єдине, що я хотіла Вас попросити. Не могли ви надіслати мені якусь доцільну літературу, яку я можу вставити в свою роботу, як практичні рекомендації, якщо людина захоче вивчати, наприклад, NLP, з чого вона може почати для ознайомлення?

Р.: Так, можу, зараз навіть, є книжка онлайн. Це така стандартна книжка, яку

я рекомендую. Тут правда старе видання, але воно в принципі досить основне, те з чого можна почати вивчення NLP, власне комплінгвістики. Там дуже хороший основний курс роботи з бібліотекою NLP Key, ту яку я згадувала, можна використовувати. З того я раджу розпочинати. Далі вже дуже багато ресурсів, які можна опрацювати.

І.: А де знаходяться всі ці бібліотеки?

Р.: Якщо ви вмієте користуватися Python, то ви просто встановлюєте їх на своє середовище і все.

Додаток Д

Список рекомендованої літератури, наданий експертами:

Barlow, M. (2013). Real Time Big data Analytics: Emerging Architecture. *O`Riley Media*.

Harrington P. (2012) *Machine Learning in action. Retrieved from* *https://www.dropbox.com/s/15doe3tjyhkiotp/machine_learning_in_action.pdf?dl=0*

Gelman, A., John B., Carlin, Hal S. Stern, D.B. Rubin. (2003). *Bayesian Data Analysis* (Second edition). CRC Press.

Shai Shalev-Shwartz, Shai Ben-David. (2014). Understanding Machine Learning: From Theory to Algorithms *Cambridge University Press.*

Tilburg University (2019). History, Evolution and Future of Big Data. *British Journal of Management*, pp. 229–251. DOI 10.1111/1467-8551.12340

James G., Witten D., Hastie T., Tibshirani R. (2017). *An Introduction to Statistical Learning.* Springer. DOI 10.1007/978-1-4614-7138-7

Hastie T., Tibshirani R., Friedman J. (2009). *The Elements of Statistical Learning.*

Cielen D., Arno D. B., Meysman, Ali M., (2016). 3. Machine learning. B *Introducing Data Science* (p. 58).

EMC Education Services. (2015). *Data Science & Big Data Analytics.*

Cai T., Zhou Y. (2016). What should sociologists know about big data? University of Macau.

Dive into Machine Learning. Retrieved from *https://github.com/hangtwenty/dive-into-machine-learning*

data-scientist-roadmap. Retrieved from *https://github.com/MrMimic/data-scientist-roadmap*

Додаток Е

Список джерел великих даних

Портал відкритих даних. https://data.gov.ua/

Zenodo. Retrieved from https://zenodo.org/

IndexDataBase. Retrieved from https://www.indexdatabase.de/

Agricultural Research Service. Retrieved from https://www.ars.usda.gov/

U.S. Department of agriculture. Retrieved from https://www.usda.gov/

Quandl. Retrieved from https://www.quandl.com/

Aminer. Retrieved from https://www.aminer.org

Internet Archive. Retrieved from https://archive.org/details/doi-urls

KDL. Retrieved from https://kdl.cs.umass.edu/display/public/DBLP

DIMACS Retrieved from. http://users.diag.uniroma1.it/challenge9/download.shtml

NBER. Retrieved from http://data.nber.org/patents/

Stanford Large Network Dataset Collection. Retrieved from http://snap.stanford.edu/data/

The ClueWeb09 Dataset. Retrieved from http://lemurproject.org/clueweb09/

CRAWDAD. Retrieved from https://crawdad.org/

CAIDA Data. Retrieved from https://www.caida.org/data/overview/

AQUASTAT. Retrieved from http://www.fao.org/aquastat/en/

National Estuarine Research Reserve System. Retrieved from http://cdmo.baruch.sc.edu/

Portal for Historical Statistics. Retrieved from http://www.historicalstatistics.org/

The New York State Education Department. Retrieved from https://data.nysed.gov/

Alberta. Retrieved from https://www.alberta.ca/index.aspx