Міністерство освіти і науки України
Харківський національний університет радіоелектроніки
Факультет Комп'ютерних наук
Кафедра Програмної інженерії

АТЕСТАЦІЙНА РОБОТА

Пояснювальна записка

рівень вищої освіти – другий (магістерський)

Дослідження сучасних засобів обробки природних мов

Виконав: студент 2 курсу, групи ІПЗм-18-1
Мачула О.В.
(прізвище, ініціали)

спеціальності 121 – Інженерія програмного забезпечення
(код і повна назва спеціальності)

Освітньо-наукової програми
(тип програми)

Інженерія програмного забезпечення
(повна назва освітньої програми)

Керівник д.т.н, проф. Четвериков Г.Г.

Допускається до захисту

Зав. кафедри, проф. _____ Дудар З.В.

2020 р.

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ

Факультет_Комп'ютерних наук_____
Кафедра_Програмної інженерії_____
Рівень вищої освіти – другий (магістерський)

Спеціальність _121 – Інженерія програмного забезпечення_____
(код і повна назва)
Тип програми освітньо-наукова програма_____
Освітня програма_Інженерія програмного забезпечення_____

ЗАТВЕРДЖУЮ:

Зав. кафедри _____
(підпис)
«_____»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

Студентові __Мачулі Олені Володимирівні_____
(прізвище, ім'я, по батькові)

1. Тема роботи __Дослідження сучасних засобів обробки природних мов
затверджена наказом університету від "___" _____20 ___ р. № _____
заповнюється вручну після отримання наказу
2. Термін подання студентом роботи до екзаменаційної комісії 18 травня 2020 р.
3. Вихідні дані до роботи обробка природних мов, методи оцінки
4. Перелік питань, що потрібно опрацювати в роботі мета роботи, аналіз проблемної галузі і постановка задачі, методи машинного навчання, проблематика обробки природної мови, загальні обробки природної мови, завдання класифікації та кластаризації

5 Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецрозділ	д.т.н, проф. Четвериков Г.Г		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка *
1	Аналіз предметної галузі		
2	Методи обробки природних мов		
3	Прототипування		
4	Підготовка пояснювальної записки		
5	Спецчастина		
6	Підготовка презентації та доповіді		
7	Попередній захист		
8	Нормоконтроль, рецензування		
9	Занесення диплома в електронний архів		
10	Допуск до захисту у зав. кафедри		
* заповнюється вручну після виконання чергового пункту			

Дата видачі завдання _____ _____ 2019 р.

Студент _____
(підпис)

Керівник роботи _____ д.т.н., проф. Четвериков Г.Г._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Звіт з науково-дослідницької практики: __ с., рис., табл., __ джерел.

МАШИННОГО НАВЧАННЯ, АНАЛІЗ ДАНИХ, ОБРОБКА ПРИРОДНОГО МОВИ, ТОКЕНІЗАЦІЯ, КЛАСТЕРИЗАЦІЯ

Метою роботи є дослідження, аналіз та реалізація методів класифікації статистичних ознак тексту, класифікація текстів, що належать різним авторам, і дослідження динаміки точності класифікації залежно від довжини текстових фрагментів.

Завдання вирішувалася в програмному середовищі Jupyter Notebook дистрибутива Anaconda, який дозволяє відразу встановити Python і необхідні бібліотеки. Для вирішення поставленого завдання використовувалися:

– Методи обробки природної мови;
– Статистичні характеристики текстів;
– Методи машинного навчання;
– Методи зниження розмірності для можливості візуалізації.

На основі отриманої динаміки зміни точності класифікації в залежно від довжин текстових фрагментів були зроблені відповідні висновки про оптимальну довжину текстів, використовуваних для навчання і тестування моделей.

MACHINE LEARNING, DATA ANALYSIS, NATURAL LANGUAGE PROCESSING, TOKENIZATION, CLUSTERIZATION

The aim of the work is to study, analyze and implement methods of classification of statistical features of the text, classification of texts belonging to different authors, and study the dynamics of classification accuracy depending on the length of text fragments.

The problem was solved in the software environment Jupyter Notebook distribution Anaconda, which allows you to immediately install Python and the necessary libraries. To solve this problem were used:

– Methods of natural language processing;
– Statistical characteristics of texts;
– Methods of machine learning;
– Dimension reduction methods for visualization.

Based on the obtained dynamics of changes in the accuracy of classification depending on the lengths of text fragments, appropriate conclusions were made about the optimal length of texts used for training and testing of models.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

ШІ – штучний інтелект

CBOW – Continuous Bag-of-Words Model

CRISP-DM – Cross-Industry Standard Process for Data Mining

DL – Deep Learning

DM – Data Mining

IE – Information extraction

GloVe – Global Vectors for Word Representation

KD – Knowledge Discovery

ML – Machine Learning

NLP – Natural Language Processing

NLTK – Natural Language Toolkit

RSS – Rich Site Summary, збагачене зведення сайту

ВСТУП

Природна обробка мови допомагає комп'ютерам спілкуватися з людьми своєю мовою та масштабувати інші мовні завдання. Наприклад, NLP дає можливість комп'ютерам читати текст, чути мовлення, інтерпретувати його, вимірювати настрої та визначати, які частини важливі.

Сьогоднішні машини можуть аналізувати більше мовних даних, ніж люди, без втоми і послідовно, неупереджено. Зважаючи на приголомшливий обсяг неструктурованих даних, що генеруються щодня, від медичних записів до соціальних медіа, автоматизація буде критично важливою для повного ефективного аналізу текстових та мовленнєвих даних

Мова людини надзвичайно складна і різноманітна. Ми виражаємо себе нескінченними способами, як усно, так і письмово. Має не лише сотні мов та діалектів, але всередині кожної мови є унікальний набір граматичних та синтаксичних правил, термінів та сленгу. Коли ми пишемо, ми часто неправильно пишемо або скорочуємо слова або пропускаємо розділові знаки. Коли ми розмовляємо, ми маємо регіональні акценти, і ми бурмочемо, заїкаємось і позичаємо умови з інших мов.

Хоча навчання під наглядом та без нагляду, а саме глибоке навчання, зараз широко використовується для моделювання людської мови, існує також потреба у синтаксичному та семантичному розумінні та досвіді в галузі, які не обов'язково присутні в цих підходах до машинного навчання. NLP важливий тим, що допомагає вирішити двозначність у мові та додає корисну числову структуру даним для багатьох додатків, розташованих нижче, таких як розпізнавання мовлення чи аналітика тексту.

Все, що ми виражаємо, містить величезну кількість інформації. Тема, яку ми обираємо, наш тон, наш підбір слів, все додає певного типу інформації, яку можна інтерпретувати та витягувати з неї цінність. Теоретично ми можемо зрозуміти і навіть передбачити поведінку людини, використовуючи цю інформацію.

Але є проблема: одна людина може генерувати сотні чи тисячі слів у декларації, кожне речення з відповідною складністю. Якщо ви хочете масштабувати та проаналізувати кілька сотень, тисяч чи мільйонів людей чи декларацій у певній географії, то ситуація не може бути керованою [1].

Дані, отримані в результаті розмов, декларацій або навіть твітів, є прикладами неструктурованих даних. Неструктуровані дані не вписуються в традиційну структуру рядків і стовпців реляційних баз даних і представляють переважну більшість даних, наявних у реальному світі. Це безладно і важко маніпулювати. Тим не менш, завдяки прогресу в таких дисциплінах, як машинне навчання, відбувається велика революція щодо цієї теми. Сьогодні мова йде не про спробу інтерпретувати текст чи мовлення на основі його ключових слів, а про розуміння значення цих слів. Таким чином можна виявити фігури мови, як іронія, або навіть провести аналіз настроїв.

Це дисципліна, яка орієнтована на взаємодію між наукою про дані та людською мовою, і масштабує багато галузей. Сьогодні NLP процвітає завдяки величезним вдосконаленням доступу до даних та збільшенню обчислювальної потужності, які дозволяють практикуючим лікарям досягти вагомих результатів у таких сферах, як охорона здоров'я, ЗМІ, фінанси та людські ресурси.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

Обробка природної мови – загальний напрямок штучного інтелекту і математичної лінгвістики. Вивчає проблеми комп'ютерного аналізу і синтезу природних мов.

Обсяг цієї глави – це теоретичні основи обробки природних мов. Через велику універсальність та величезну кількість різних застосувань у NLP, в цій главі буде зосереджено лише тематика, яка є важливою для цієї роботи. Розділ розпочинається з пояснення виявлення знань та міжгалузевого стандартного процесу для видобутку даних. Він продовжується розслідуванням NLP загалом і обговорює технології та методи, які мають вирішальне значення для виконання цієї дипломної роботи. Внаслідок сучасні тенденції в NLP, таких як нейронна мережа прямого поширення та рекурентна нейронна мереж.

Для розуміння більшої картини роботи я коротко введу терміни процес пошуку корисних знань в базах даних та глибинний аналіз даних. У літературі видобуток даних та KD у базах даних часто трактуються як синоніми, але насправді, майнінг даних – це під задача KD-процесу, схожа на описану CRISP-DM у наступному розділі. DM може розглядатися як фаза моделювання, яка витягує створення та зразки підготовлених даних з різними підходами.

Міжгалузевий стандартний процес для дослідження даними був розроблений в 2000 р. і застосовується до багатьох різних проблем з видобутком даних [2]. В деталях, процес складається з шести основних етапів – розуміння бізнесу, розуміння даних, підготовка даних, моделювання, оцінка та розробка. Крім того, весь процес є ітераційним, тобто кожен крок буде оброблятися кілька разів.

Розуміння бізнесу – головне завдання на початку процесу обміну даними. З точки зору бізнесу, весь проект слід розуміти та аналізувати. Крім того, ця фаза намагається перетворити всю проблему в перспективу обміну даними, яка зосереджена на похідних цілях бізнесу.

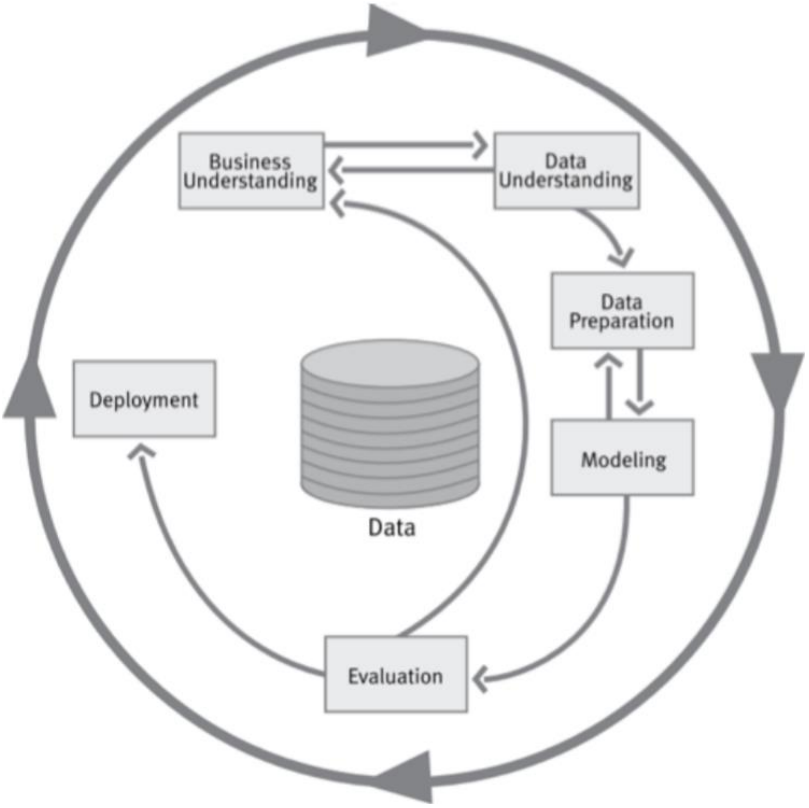

Рисунок 1.1 – Міжгалузевий стандартний процес для дослідження даними

Збір даних – це початковий крок у розумінні даних. Цей крок служить для забезпечення початкові уявлення про дані, формують ранні гіпотези та виявляють проблеми в даних. Таким чином можна переробити всю проблематику і повернутися до справи розуміння. Згодом акцентується увага на підготовці даних, де різні методи застосовуються для отримання начального набору даних, які можуть бути використані як вхідні дані при моделюванні фази. Відповідні завдання для підготовки даних в NLP описані в наступному розділі. Крім того, попередньо обробляючи дані зазвичай розбиваються на менші підмножини перед моделюванням. Один називається навчальним набором, на якому тренується модель, другий – тестовим набором і він використовується для оцінки. Іноді третій набір, який називається набором перевірки, надає дані для навчання і параметри DM-моделі.

Моделювання – це етап вибору та застосування різних моделей до підготовлених даних. Низка методів для кожної окремої задачі вибору даних. Часто необхідно повернутися до етапу підготовки даних, оскільки деякі моделі

вимагають різного рівня форми введення [3]. Перед розробкою, де модель передається у виробництво. Необхідно оцінити застосовані моделі за допомогою показників, орієнтованих на бізнес-цілі. Тому крок оцінки вимірює ефективність усіх попередніх етапів за допомогою дотримання цілей бізнесу.

1.1. Обробка природних мов

Обробка природних мов є основною дослідницькою сферою цієї дипломної роботи. Це пов’язано з іншим сферами, таким як штучний інтелект або машинне навчання. Різні методи попередньої обробки функцій, наприклад, видалення стоп слова та векторна модель. У цьому розділі також розглядаються показники оцінки та обговорюється класична машина модель навчання.

В розділі розміщено теми – штучний інтелект, обробка природних мов, машинне навчання та глибоке навчання в одному контекст. ШІ – це дуже широкий термін і є способом описати системи, які здатні «думати». У літературі існує багато різних пояснень, як визначити цю тему. Інтерпретація буде прийняти, чому включає NLP і ставить його по відношенню до ШІ. ШІ складається з чотирьох основних частини, які є машинним навчанням, аргументація, плануванням та NLP. Обґрунтування дозволяє машині надавати пропозиції на основі даних, тоді як планування дає змогу системам самостійно діяти при інтерпретації даних.

У ньому є багато різних застосувань, які відносяться до неструктурованої природної мови. Наприклад, сферами його застосування є машинний переклад, розпізнавання мови, діалогові системи, розпізнавання іменованих об'єктів, пошук інформації та класифікація тексту . Таким чином, домен NLP охоплює всі взаємодії між комп'ютером і людиною, шляхом використання писемної чи розмовної природної мови.

Ця дослідницька та прикладна робота, що стосується маніпулювання та розуміння природних мов. Обробка людської мови заснована на розумінні

наміченого значення повідомлення, яке є важким навіть для людей, наприклад, коли використовується іронія. Усі компоненти природної мови, таких як фонетика, фонологія, морфологія, синтаксис, семантика та прагматики, повинні враховуватися, щоб отримати повне розуміння повідомлення.

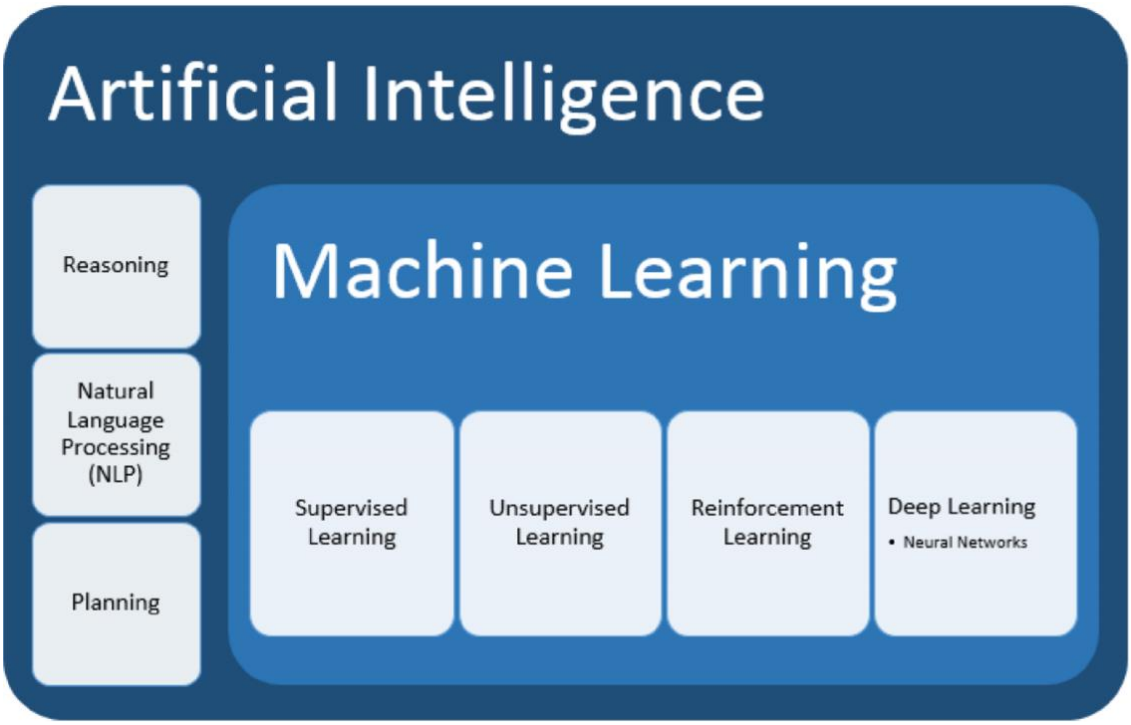

Рисунок 1.2 – Штучний інтелект містить машинне навчання та NLP

Фонетика – це про акустичні властивості звуку, що видається голосовим трактом людини. Вивчає, як звуки фізично побудовані, наприклад, з мовою або губами. Звук конкретної людської мови вивчається фонологією. Наприклад, англійська мова має 45 відмінних звуків, званих фонемами. Фонетика та фонологія особливо важливими аспектами розпізнавання мовлення при перетворенні звуків у реальні слова, які можна обробити комп’ютером.

Морфологія стосується значення та архітектури слів. Слово-будування і лематизація, які описані нижче, базуються на цьому компоненті шляхом перетворення слова.

Впорядкованість слів та побудова граматично правильних речень досліджено синтаксис. Навпаки, семантика вивчає значення побудованих речень шляхом використання синтаксичних та морфологічних словоформ.

Для отримання загального значення повідомлення, прагматика використовує контекст ситуації [4]. Наприклад, припустимо, хтось запитує: «Чи могли б ви передати сіль?». Дано, у контексті ситуації питання фактично є проханням передати сіль, а не запитання, чи хтось може це зробити. Тому комп’ютер потрібно брати всі частини природної мови врахувати, щоб нею користуватися.

Одне відоме визначення машинного навчання засноване на ідеї досвіду та ілюструє навчальну частину в ML: «Комп'ютерна програма, як кажуть, вивчає досвід щодо деяких клас завдань і міра ефективності, якщо її виконання при завданнях в, які вимірюється, покращується з досвідом.».

Наступний приклад повинен допомогти уточнити це поняття. Припустимо, є дощсистема прогнозування, яка прогнозує, буде сьогодні дощ чи ні. Система – бінарний класифікатор і його продуктивність буде вимірюватися точністю. Дізнається з часу та історичних погодних даних передбачити прогноз правильний результат.

Крім того, ML є одним із центральних предметів у ШІ та розбивається на чотири підкатегорії: контрольоване навчання, непідконтрольне навчання, глибоке навчання та навчання з підкріпленням. Контрольоване та непідконтрольне навчання – це два основних типи проблем пошуку даних. Різниця між цими завданнями полягає в структурі даних про навчання. У контрольованому наборі дані, цільова змінна відома і може використовуватися для модельного навчання. Непідконтрольне завдання леми не мають відомого результату, і рішення часто базується на подібності екземпляра, візерунка та групи.

Навчання підкріпленню схоже на наглядове навчання, але не вчиться на набору даних. Її часто використовують в ігрових іграх або на самостійних керування автомобілями. Навчання зміцненню алгоритми використовують пробні та помилкові знання для даної задачі. Взаємодія із оточення та використовування зворотного зв’язку для покращення свого досвіду.

На відміну від трьох згаданих вище тем, Deep Learning – це метод їх вирішення, дозволяючи комп'ютеру робити складні шаблони з більш простих. Deep neural networks є основними частинами DL [5]. Згідно з вищенаведеним

визначенням, сказати, використовується DL чи, важко ні, тому що складна картина може бути також нейронною мережею з одним прихованим багато нейронів. Мережа не повинна бути глибокою у кожному вимірі.

Терміни штучний інтелект, обробка природної мови, машинне навчання та DL не слід розуміти лише самі по собі. Межі між темами не суворі. Зокрема, ML широко використовується в NLP для вирішення різних типів проблем.

Вибір функцій та попередня обробка є важливими завданнями в галузі штучного інтелекту та в основному представляють етап підготовки даних в CRISP-DM. Особливо в NLP це завдання має величезний вплив на успіх аналізу тексту. В основному це викликані неструктурованим та довільним характером текстових даних. Крім того, машини потребують структури та числових даних. Кілька підходів до цієї задачі трансформації, наприклад, існують вбудовані слова або модель векторного простору. Обсяг цього розділу лежить на теоретичне підґрунтя різних методів попередньої обробки та вибору функцій. Цей розділ супроводжуватиметься англійською фразою „the best fox are run“, як приклад для ілюстрації застосування попередньої обробки.

Для обробки писемної природної мови неминуче розбивати тексти на менші одиниці, які називаються лексемами. Комп'ютери повинні відрізняти окремі об'єкти тексту та для їх створення використовується токенізація. Зазвичай лексеми являють собою прості слова, які найменші самостійні одиниці природної мови. Крім того, лексеми можуть складатися з ідіом чи дефісів, наприклад створених користувачем. Токенізація розбиває текстові тексти на короткі текстові об'єкти і це найперше завдання у будь-якому циклі попередньої обробки тексту. Окрім розділу на невеликі одиниці, цілі речення також можуть бути результатом токенізації. Простий токенізатор слів може бути реалізований у багатьох мовах шляхом розбиття тексту при появі символів. Цей простий базовий підхід має пару недоліки, через відсутність ідентифікуючих слів, які семантично належать разом. Однак простий токенізатор розділяє фразу, яку було введено вище, на наступні п’ять лексем: “the” “best” “fox” “is” “running”.

За допомогою лексем можна створити так звані n-грами, які позначають набір лексем із символом довжина.

Дуже важливим підходом до зменшення величезного простору вхідного простору в NLP є стоп слова. Більшість мов мають специфічні слова, які з’являються частіше, ніж інші не включають багато інформації про зміст тексту, наприклад, допоміжні дієслова чи статті. Завдяки цьому часто має сенс виключити ці так звані стоп слова в подальшому аналіз. Англійською такими словами можуть бути “the”, “a” або “an”, а для німецької мови типова стоп слова – артикули “die”,“der” та “das”. Усунення можна зробити за допомогою перевірки слова проти стандартизованого списку стоп слів. Ці списки доступні в літературі та часто реалізуються в різних програмних пакетах. У нашому прикладі “the” та “is” усуваються. Стоп слова слід використовувати обережно, особливо в аналізу тональності тексту, який намагається передбачити позитивний чи негативний намір тексту [6].

Окрім усунення слів, стемінг – корисна техніка для зіставлення слів до їх слова додатково зменшують вхідний розмір. Це допомагає витягнути реальне значення тексту і робить неструктуровані дані краще доступними для машини. Перший стемінг алгоритм, заснований на вилученні найдовших суфіксів та написання виключень з написання, був розроблений у 1968 році. На сьогоднішній день алгоритм визначення носія – це найсучасніший підхід, який супроводжує суфікси слів, щоб зберегти слово стовбур . Хоча цей метод добре працює в з англійською мовою, є деякі недоліки для німецької мови через те, що німецькі слова, як правило, не будуються шляхом додавання суфіксів. Однак є німецький еквівалент заснований на ідеї Портера та мови Snowball.

Лематизація – це процес зіставлення кожного слова в тексті на їх тип словника або призначення похідної структури. Дієслова перетворюються на їх інфінитивну форму, іменник реконструюється до його однини і прислівники або прикметники передбачають їх позитивну форму. Метод заснований на морфологічному аналізі і часто використовується словниками, наприклад WordNet, де можна знайти лему кожної модифікованої форми слова. Цей крок попередньої

обробки схожий на стримування та зменшує вхідний простір шляхом зіставлення різних форм слів до їх загального представлення. З тих пір лематизація підтримується словниковими записами.

Крім того, попередньо обробляючи самі слова, їхні уявлення потрібно змінити у форматі для машинного читання. Тим часом було декілька різних підходів розроблених для перетворення текстів у різні види числових уявлень. Деякі з них представляють лише статистику слова, наприклад, one-hot-encoding та інші формати також включайте контекст слова, наприклад, word2vec.

Векторна модель – це підхід, який перетворює текст в один вектор заснований на однокольоровому кодуванні слів. Враховуючи набір текстових документів , можливо створити словниковий запас довжиною N. Однокольорово закодований вектор словапредставляє слово на 1 у відповідному словниковому записі [7]. Наприклад, якщо термін “fox” – це i-унікальне слово в корпусі, вектор має довжину N та 1 у i позиція, всі інші записи 0.

Модель векторного простору поширює цю модель на документи. Функція Ψ відображає будь-який документ d до його векторного представлення простору. Виразні слова, які також називаються терміни, у лексиці представлені $t_1 , \ldots , t_n$.

$$\Psi: d \mapsto \Psi(d) = \left(tf(t_1, d), tf(t_2, d), \ldots, tf(t_n, d)\right) \in R^N \qquad (1.1)$$

Ψ підраховує кількість кожного терміна (tf) у словниковому запасі на документ. Отже, вектор документа може мати більше одного запису, який не дорівнює 0. Функція $tf(t_i, d)$ визначає, наскільки часто i словникове слово з'являється в документі d. Окрім простого використання терміна частота (tf) , можна також дати кожну конкретнуслово зважування, відповідно до його відносного появи в корпусі. Це можна зробити використовуючи термін частоти, розділений на зворотну частоту документа $(tf - idf)$, що дає менше значення для загальних слів у корпусі. Щоб обчислити $tf - idf$ для терміна в документі,в першу чергу слід визначити нормовану частоту. Відносний нормований термін частота (nef_{rel}) визначається як:

$$ntf_{rel}\,(t_i, d) = \frac{tf\ (t_i, d)}{\sum_{t_m \in d} tf(t_m, d)} \tag{1.2}$$

Він обчислює частотну частину i доданку t_n в d і ділить це на суму всі інші частотні частоти в d. Максимальна нормалізована частотна частота (ntf_{max}) ставить найвища частотна частота в d в знаменнику:

$$ntf_{max}\,(t_i, d) = \frac{tf\ (t_i, d)}{max_m tf(t_m, d)} \tag{1.3}$$

Обидва рівняння можуть бути використані для отримання коригованого значення терміна протягом конкретного документу d. Для включення слова релевантність щодо цілого корпусу, необхідно обчислити зворотну частоту документа

$$idf(t_i) = \log\left(\frac{|D|}{|d: t_i \in D|}\right) \tag{1.4}$$

$|D|$ – загальна кількість документі, а знаменник являє собою виникнення терміна t_i у всіх документах. Для частих доданків $|d: t_i \in D|$стає великим, що призводить до частки ближче до 1 [8]. Повернення частоти документа на логарифм, загальні слова мають меншу вагу, тому що їх менше відмінність, ніж рідші слова. Фактична вага w для кожного слова в документі обчислюється добутком нормалізована частотна частота та зворотна частота документа:

$$w(t_i, d) = id\ f(t_i) \cdot ntf(t_i, d) \tag{1.5}$$

Використовуючи векторну модель, розрідженість векторів документів може бути одна. Основна проблема пов'язана з тим, що N-тривалість (d) позицій

дорівнює 0. Ще одна проблема полягає в тому, що відстань між двома різними векторами документів дуже мала. Тому розрізнити документи, особливо, непросто якщо мова йде про групування подібних документів, наприклад, при кластеризації.

1.2 Методи оцінки

Етап оцінювання є однією з основних частин будь-якого проекту вилучення даних і є успішним крок моделювання. У цьому розділі представлені різні методи оцінки та вимірювання виконання різних підходів ML. Особливо представлені показники, до яких оцінка зауважених навчальних завдань [9]. Він починається з визначення матриці плутанини і закінчується описом заходів оцінювання, таких як точність та відкликання.

Матриця плутанини – це метод представлення результатів контрольованого навчального завдання і використовуються в задачах класифікації, таких як приклад прогнозування дощу. Система передбачає, буде дощ чи ні, що може бути реалізовано двома класами «дощ» та «сухий». Матриця плутанини порівнює справжні мітки екземплярів з передбачувані заняття. Сума діагоналі матриці є сумою загальної істини прогнози

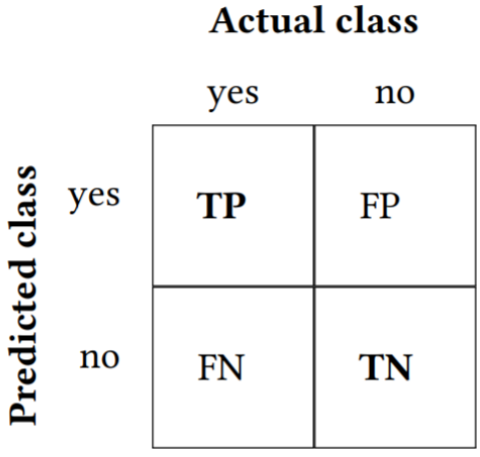

Рисунок 1.3 – Матриця плутанини

Матриця плутанини містить значення для True Positive (TP) , True Negatives (TN), False Positives (FP) та False Negatives (FN). Матриця проілюстрована для бінарний клас з двома результатами (так, ні).У ситуації з декількома маркуваннями матрицю плутанини легко розширити, додавши стовпчик і рядок на клас. Зразкова матриця з двійковим класом проілюстрована на рисунку 1.3, де TP означає кількість справжніх позитивних, TN – для справжніх негативів, FP – для помилкових позитивів і FN – для помилкових негативів [10]. Застосовано до прикладу прогнозу дощу, TP було б, якби система спрогнозувала клас «дощ», а в цей день погода була дощовою. А TN вказувало б на те, що передбачувалося «сухе», що було дійсно на той день, тоді як FP виражає, що система передбачила «суху». FN є протилежністю FP.

Щоб отримати огляд продуктивності конкретного алгоритму, точність (accuracy) – гарний вибір може бути визначена для задачі класифікації. Точність обчислюється сумою діагоналі, поділену на суму всіх записів у матриці плутанини

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.6}$$

Recall – це міра, яка обчислює відсоток відповідних екземплярів алгоритму вибрав. Таким чином, він оцінює, скільки TP знайдено у фактичного класу.

$$Recall = \frac{TP}{TP + FN} \tag{1.7}$$

Precision – ще одна міра оцінки. Він обчислює відсоток TP у наборів якій модель класифікувала всі випадки як позитивні.

$$Precision = \frac{TP}{TP + FP} \tag{1.8}$$

F-міра являє собою гармонічне середнє значення між відкликанням та точністю. Зважування може регулюватися параметром β. Це робить F_β прийнятним для різних завдань з видобутку даних,наприклад, у пошуковій системі, де виклик може бути важливішим за точність.

$$F_\beta = \frac{(1+\beta^2)\cdot precision \cdot recall}{\beta^2 \cdot precision \cdot recall} \tag{1.9}$$

Середньоквадратична помилка кореня (RMSE) використовується для оцінки заданого прогнозу для набору екземплярів. Обчислення квадрату розносного між істинним результатом і прогнозом.

$$RMSE(y_i, \widehat{y_i}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(|\widehat{y_i} - y_i|)^2} \tag{1.10}$$

Середнє абсолютне відхилення (MAD) аналогічно використовується як RMSE. Він обчислює абсолют відмінність від набору екземплярів до їх середнього значення.

$$MAD(y) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \overline{y_i}| \tag{1.11}$$

1.3 Виявлення проблем та актуалізація рішень

Основною проблемою обробки природної мови є мовна неоднозначність. Існують різні види неоднозначності: синтаксична, смислова неоднозначність, відмінкова неоднозначність і т. д.

Центральна проблема, як для загальної, так і для прикладної обробки природної мови – дозвіл такого роду неоднозначності – вирішується за допомогою перекладу зовнішнього подання природної мови в якусь внутрішню структуру. Для загальної обробки природної мови таке перетворення вимагає набору знань про реальний світ.

Прикладні системи обробки природної мови мають перевагу перед загальними, тому що працюють у вузьких предметних областях.

Проте, створення систем, що мають можливість спілкування на природної мови в широких областях, можливо, хоча поки результати далекі від задовільних.

У міру розвитку комп'ютерних систем стає все більш очевидним, що використання цих систем набагато розшириться, якщо стане можливим використання людської мови при роботі безпосередньо з комп'ютером, і зокрема стане можливим управління машиною звичайним голосом в реальному часі, а також введення і виведення інформації у вигляді звичайної людської мови.

Існуючі технології розпізнавання мови не мають поки достатніх можливостей для їх широкого використання, але на даному етапі досліджень проводиться інтенсивний пошук можливостей вживання коротких багатозначних слів для полегшення розуміння. Розпізнавання мови в даний час знайшло реальне застосування в житті, мабуть, тільки в тих випадках, коли використовується словник скорочений до 10 знаків, наприклад при обробці номерів кредитних карт і інших кодів доступу в базуються на комп'ютерах системах, що обробляють передані по телефону дані. Так що нагальна задача – розпізнавання, 20 тисяч слів природної мови залишається поки недосяжною. Ці можливості поки недоступні для широкого комерційного використання. Однак ряд компаній самотужки намагається використовувати вже існуючі в даній галузі науки знання.

Для успішного розпізнавання мови слід вирішити такі завдання:

– обробку словника, фонемний склад;
– обробку синтаксису;
– скорочення мови, включаючи можливе використання жорстких сценаріїв;

- вибір диктора, включаючи вік, стать, рідна мова і діалект, тренування дикторів;
- вибір особливого виду мікрофона;
- умови роботи системи і отримання результату із зазначенням помилок.

Існуючі сьогодні системи розпізнавання мови ґрунтуються на зборі всієї доступної (часом навіть надлишкової) інформації, необхідної для розпізнавання слів. Дослідники вважають, що таким чином завдання розпізнавання зразка мови, заснована на якості сигналу, підданого змін, буде достатньою для розпізнавання, але, тим не менш, в даний час навіть при розпізнаванні невеликих повідомлень нормальної мови, поки неможливо після отримання різноманітних реальних сигналів здійснити пряму трансформацію в лінгвістичні символи, що є бажаним результатом.

1.4 Розуміння при автоматичній обробці тексту

Граничної завданням штучного інтелекту є розуміння природної мови. При цьому в якості введення і виведення може використовуватися як мова, так і текст на одному з природних мов, представлений в письмовій формі. Відповідно, обробка текстів на природній мові розпадається на дві великі завдання: обробка мови і обробка тексту.

Без вирішення завдання розуміння тексту обробка мови має прикладний характер, тому основний акцент все ж ставиться на розумінні тексту. Тому останнім часом з'являється все більше робіт, де замість звичного терміну «автоматична обробка тексту» використовується термін «розуміння». Основи машинного розуміння тексту на сучасному рівні викладені, наприклад, в монографії Є. Овчинниковой.

Завдання машинного розуміння тексту в даний час поділяється на дві області:

- автоматичну обробку тексту , в процесі якої суцільний потік символів, що надходить в машину, набуває структуру тексту, побудованого відповідно до законів природної мови;
- подання знань (Knowledge Representation, KR), тобто відображення вхідних текстової інформації на природній мові в формі, придатній для подальшої машинної обробки.

Після вирішення проблеми машинного розуміння настає черга генерування вихідної інформації на природній мові. Останнє завдання, знову ж таки, носить вторинний характер, тому що вона обов'язкова аж ніяк не у всіх областях: наприклад, при формулювання завдання на природній мові в робототехніці на виході може бути конкретна дія.

До того ж стан справ в області NLG на сьогоднішній день йде набагато краще: завдання акустичного виведення тексту незрівнянно легше завдань розуміння.

Задача вилучення інформації – розглядає питання ідентифікації певних сутностей і відносин в неструктурованих даних, особливо в текстових документах. Таким чином, вилучення інформації з тексту стає ключовим компонентом в процесі інтеграції текстових даних, і може розглядатися в якості узагальнюючого терміна для багатьох цікавих завдань, таких як розпізнавання імен сутностей, аналіз тональності тексту або витяг знань.

Витяг подій з неструктурованих даних, таких як повідомлення, може бути корисно для подальшого практичного застосування. Наприклад, якщо система IE в змозі визначити подія, то це може підвищити продуктивність персоналізованих інформаційного систем, так як новинне повідомлення може бути обрано більш точно, в залежності від уподобань користувача.

Витяг інформації є широким полем досліджень і тісно пов'язано з декількома дисциплінами:

- обробка природної мови,
- інтелектуальний аналіз тексту,
- машинне навчання.

Завдання вилучення інформації не отримала такого широкої уваги, як інформаційний пошук, і часто змішується з останнім. Завдання інформаційного пошуку тексту полягає в тому, щоб вибрати з набору текстових документів підмножина, яке має відношення до конкретного запиту, на основі пошуку по ключовим словам. Процес IR зазвичай повертає ранжируваних список документів, де ранг відповідає балу релевантності, який система привласнила документу у відповідь на запит. Однак ранжируваних список документів не надає докладну інформацію про зміст цих документів. Мета IE не ранжувати або вибрати документи, а витягти з документів важливі характеристики про попередньо визначених типах подій, сутності або відносинах. Підводячи підсумок, IE прагне перетворювати колекції текстових даних в форму, яка полегшує пошук і виявлення знань в таких колекціях.

Системи IE в цілому більш важкі і наукомістких для побудови, між системи IR. Однак методи IE і IR можуть розглядатися як взаємодії доповнюють і потенційно можуть бути об'єднані різними способами. IR часто використовується в IE для попередньої фільтрації дуже великої колекції документів і зведенні її до керованого множиною, до якого можуть бути застосовані методи IE. Як альтернативи, IE може використовуватися як субкомпоненту IR-системи для ідентифікації структур або для інтелектуального індексування документів.

Розглянемо як приклад вилучення інформації про подію з пропозиції, ми зацікавлені у визначенні часу проведення, виявленні основних учасників цього заходу і його місцезнаходження: «У червні в Києві пройшов фестиваль квітів, в якому взяли участь школярі міста.»

У таблиці 1.1 приклад структурованої інформації, яка напівучена із зазначеного вище пропозиції. Процес вилучення такої структурованої інформації включає в себе ідентифікацію деяких структур, таких як словосполучення, що позначають особу або групу осіб, географічні посилання і знаходження семантичних відносин між ними.

Завдання вилучення інформації полягає у виявленні заздалегідь заданого класу сутностей, відносин і подій в текстах природною мовою, а також вилучення відповідних властивостей ідентифікованих сутностей, відносин або подій.

Таблиця 1.1 – Приклад структурованої інформації

Час проведення	червень
Місце проведення	Київ
Учасники	школярі

Інформація попередньо яку видобувають задається в призначених для користувача структурах, званих шаблонами, кожен з яких складається з декількох слотів (або атрибутів) які повинні бути створені системою ІЕ при обробці тексту. Зазвичай слот заповнюють: рядки з тексту, одне з декількох наперед значень або посилання на раніше створений шаблон об'єкта. Система ІЕ створює структуроване уявлення обраної інформації, витягнутої з аналізованого тексту.

1.5. Архітектура системи ІЕ

Типова система ІЕ включає в себе кілька етапів. Попередня обробка вхідних текстів. Текст часто складається з неструктурованих, «сирих» текстів природною мовою. Більшу частину релевантної інформації можна виділити за допомогою закономірностей, які виявлені в лінгвістичних властивості текстів. Таким чином, завдяки лінгвістичного аналізу, можна визначити важливі особливості тексту. Для вилучення інформації використовуються наступні лінгвістичні компоненти:

Токенізація. Текст – це послідовність символів. мета токенізації полягає в тому, щоб визначити елементарні частини природної мови, такі як: слова, розділові знаки – такі елементами називаються токенами. Отримана послідовність значущих токенов є основою для подальшої лінгвістичної і будь-якої текстової обробки.

Поділ речень. Речення є одним з найважливіших елементів природної мови для структурованого подання письмової змісту. Речення є найменшими одиницями для вираження завершених думок або подій. Тому правильне визначення меж речення має важливе значення для багатьох підходів ІЕ. Дане завдання було б тривіальної, якби розділові знаки не використовувалися неоднозначно. Для синтаксичного аналізу необхідно правильне уявлення тексту в вигляді послідовності речень.

Морфологічний аналіз. Деякі факти зазвичай виражаються певними частинами мови (наприклад, імена – іменниками). Визначення частин мови токенов називається POS-тегування. Системи на основі машинного навчання можуть використовувати POS-теги в якості класифікаційних ознак, системи на основі правил, як елементів правил вилучення.

NER, CO. Завдання розпізнавання іменованих сутностей і дохволення кореференції іменованих сутностей є одними з найбільш часто вирішуються системами ІЕ. Деякі підходи використовують простий пошук в зумовлених списках, деякі використовують приховані Марківські моделі для ідентифікації іменованих сутностей і їх типу. Завдання СО полягає в знаходженні декількох посилань на один і той же об'єкту тексті. Це особливо важливо, оскільки відповідне утримування може бути виражено займенниками і позначеннями. Обидва завдання вимагають глибокого семантичного аналізу і не так надійні, як інші лінгвістичні компоненти.

Для систем, заснованих на онтологіях і правилах, лінгвістична попередня обробка є одним з базових елементів. для систем, заснованих на машинному навчанні попередня обробка текстів є необов'язковою, але може мати серйозний вплив на якість вилучення.

Застосування моделі вилучення інформації. Діапазон при трансформаційних змін сучасних систем ІЕ повинен бути максимально широким. особливості конкретної області не можуть бути закріплені в системі, оскільки зусилля з адаптації до інших областей занадто великі [11]. Сучасні системи використовують компонент навчання для зменшення залежності від конкретних областей і

зменшення обсягу ресурсів, що надаються людиною. Модель вилучення визначається відповідно до застосовуваного підходом, і її параметри оптимізуються за допомогою процедури навчання. Підходи на основі машинного навчання вивчають, наприклад, відсутність класифікаційні ознаки, ймовірності. підходи основані на правилах вивчають набір правил вилучення. Підходи на основі онтології вивчають структури для доповнення та інтерпретації своїх знань для подальшого вилучення. Завдання полягає в тому, щоб знайти модель вилучення, яка дозволяє вивчити всі відповідні параметри області.

Рис 1.4 – Архітектура IE системи

З огляду на проблеми і складність IE, контрольоване навчання представляється найбільш підходящим і є широко використовуваною технікою навчання. Більшість підходів воліють анотовані навчальні корпусу, хоча деякі покладаються на людський нагляд на етапі навчання. Для оцінки якості підходу навчальний текстовий корпус створюються шляхом анотування фрагментів текстів

з релевантним вмістом, розділених на дві частини. Одна частина, навчальний набір, використовується для навчання, а інша, тестовий набір, використовується для перевірки здатності моделі правильно залучати нову інформацію, на якій вона не була навчена. результати тесту також можуть бути використані для поліпшення роботи моделі потягу.

Деякі підходи дозволяють подальше удосконалення поділи вилучення на основі зворотного зв'язку від людини про витягах у час застосування. Нові оцінені вилучення можуть бути включені як нові навчальні екземпляри, і модель може бути перенавчитися. Компонент навчання має вирішальне значення для системи ІЕ, оскільки він включає в себе алгоритми ідентифікації відповідних частин текста і передачі їх відповідно до цільової структурою.

Піст обробка вивідних текстів. Основною мотивацією для ІЕ є структуроване подання інформації, що дозволяє виконувати формальні запити і автоматичну обробку. Однією з множин структурування витягнутих даних є моделювання цільової структури як відносини бази даних. Після того, як відповідна інформація була знайдена за допомогою застосування моделі вилучення, ідентифікованим фрагментами тексту присвоюються відповідні атрибути цільової структури. Вони можуть бути нормалізовано відповідно до очікуваного форматом. Деякі ідентифіковані факти можуть з'являтися в тексті більш ніж один раз або вже знаходитися в базі даних. В цьому випадку різні екземпляри можуть бути об'єднані. Нарешті, ідентифікація, нормалізована і уніфікована інформація зберігається в відповідній базі даних.

Візуалізація архітектури типової системи вилучення інформації з тексту представлена на рисунок 1.4.

2 ОПИС ТЕОРЕТИЧНИХ ДОСЛІДЖЕНЬ

У цій главі наведено огляд найсучасніших технологій та підходів у NLP. Обсяг визначається першим дослідницьким питанням: «Який найсучасніший NLP підходи існують для аналізу текстів опису виробів, створених користувачем?». Дослідження показали, що аналіз текстів лише для прогнозування на результат бізнесу. Цей напрямок досліджень фокусується на прогнозуванні цін на фондових біржах за допомогою NLP. Дослідники аналізують тексти, такі як тексти соціальних медіа, фінансові звіти або новинні тексти. Ще один подібний дослідницький аналіз – це аналіз, створений користувачем огляди товарів, що задає акцент на тому, що впливає на рішення покупців користувачів. Обидві галузі дослідження відрізняються від мого підходу в плані структури тексту чи цілі. Таким чином, зупиняючись на більш загальних методах, які використовуються для прогнозування мети тексту.

Окрім методів моделювання, дуже важливими є також функції, що представляють вхідні дані важливо для будь-якого завдання з пошуку даних. Особливо при роботі з неструктурованими даними, наприклад, тексти природничою мовою, важливе значення має породження та трансформація особливостей за досягнення хороших результатів [12]. Таким чином, глава розпочинається з пояснення таких методів вибудовування слів і продовжується глибокими контекстуалізованими поданнями слів.

2.1 Порівняння існуючих видів бібліотек

Існують різні програмні інструменти для побудови систем вилучення інформації. Розглянемо найбільш популярні з них:

- Stanford CoreNLP – програмний засіб, який продається для створення представлених аналізів текстів на естетичній мові;
- NLTK – це важлива бібліотека, яка підтримує такі завдання, як класифікація, стемінг, маркування, синтаксичний аналіз та семантичне розширення в Python;
- spaCy відносно молода бібліотека, передбачена для виробничого використання. Почитайте, що вона отримала доступ до інших NLP-бібліотек Python, таких як NLTK. spaCy пропонує найсильніший синтаксичний парсер, який сьогодні знаходиться на ринку. Крім того, за допомогою інструментарію написаний на мові Cython, він також дуже важливий і ефективний.

Окрім опису вирішальної теорії, я коротко опишу технології, які застосований у виконанні моєї дипломної роботи. Різні пакети Python підтримують частину кодування, особливо PyTorch та spaCy

spaCy – бібліотека з відкритим кодом Python для завдань NLP. Його архітектура ідеально підходить для використання в розробці. Крім того, постачає реалізації для попередньої обробки тексту, глибокого вивчення та інші завдання в галузі NLP. У цій роботі використовується в основному для попередньої обробки тексту, такі як токенізація, лематизація та стоп слова. Методи засновані на моделі німецької мови, яка навчалася у Вікіпедії та корпусі TIGER. Крім того, spaCy підтримує більше 31 мов і має заздалегідь підготовлені слова використовуючи підхід word2vec.

Prodigy та Thinc – це дві платформи, що розширюють spaCy. Prodigy – це веб інструмент для створення швидкого та різного обсягу навчальних даних. Починає тренувати модель на невеликий набір даних і вносить пропозиції для нових примірників. Користувач може тегувати, чи модель була правильною чи ні. Thinc є бібліотекою глибокого навчання spaCy і підтримує швидку сучасні моделі в NLP, засновані на архітектурі «embed, encode, attend, predict» .

PyTorch – це платформа, яка підтримує глибоке навчання та надає бібліотекам легко створювати штучну нейронну мережу. Використовується для побудови лінійного багатошарового перцептора і довготривалої пам'яті.

Однією з головних переваг PyTorch є реалізовані моделі, функції та алгоритми навчання. PyTorch заснований на так званих тензорах, які представляють багатовимірні матриці. Всі обчислення виконуються тензорними операціями. Крім того, він реалізував підтримку графічних процесорів , що полегшує навчання та аналіз моделі нейронної мережі на графічних процесорах. Завдяки багатьом тензорним операціям, які будуть виконуватися протягом навчання має сенс використовувати їх, оскільки вони призначені для виконання матричних обчислень.

Stanford CoreNLP являє собою набір інструментів для аналізу природної мови людини. За допомогою даних засобів можливе знаходження основи слова, частини мови, дат, часових проміжків і числових величин. Також інструменти дозволяють розмітити структуру пропозиції в термінах фраз і синтаксичних залежностей, вказати, які фрази відносяться до тих чи інших сутностей. Можливе здійснення аналізу тональності тексту, витяг імен та відносин. На рисунку 2.1 представлені приклади розв'язання задач розпізнавання іменованих сутностей і дозволу кореференції.

Stanford CoreNLP включає в себе різні модулі:

- Stanford Named Entity Recognizer – модуль, для розпізнавання іменованих сутностей;
- Stanford Relation Extractor і Stanford OpenIE – модулі для вилучення відносин і фактів з тексту;
- Stanford Pattern-based Information Extraction and Diagnostics – модуль для ітеративного побудови шаблонів для вирішення завдання IE.

Використовувані модулі написані на мові Java. Кожне з засобів є для розробників через API і CLI.

Natural Language Toolkit – це набір бібліотек і програм для символьної та статистичної обробки природних мов, написаної мовою програмування Python. Її розробили Стівен Берд й Едвард Лопер.

NLTK визначає інфраструктуру, яку можна використати для побудови програм NLP у Python. Вона надає базові класи для представлення даних, що мають відношення до обробки природної мови; стандартні інтерфейси для виконання таких завдань: анотування частин мови, синтаксичний розбір і класифікація тексту; і стандартні реалізації для кожного завдання, які можуть бути об'єднані для вирішення складних завдань.

2.2 Векторне представлення слів

Векторне представлення слів – це загальний підхід до відображення великих розмірних векторів слів, таких як, однокольорові кодовані вектори, в низькомірне зображення. Основна ідея була вперше згадано Йошуа Бенжио у 2001 році і був мотивований подолати прокляття розмірності. Векторне представлення слів обчислюються для набору текстів зі словниковим складом розмір N. Для ілюстрації, припустимо, ми хочемо створити вбудови для єдиний документ: "the best fox is running". У реченні є виразні слова, які веде до $N = 5$. Слово w_i в цій лексиці $i = 1, \ldots, N$ подається вектором із заданим розміром. Якщо використовується вбудований розмір (ED) 4, зберігає вбудовувані слова для всіх $N = 5$ слів у документі. E-вбудований $N \times ED$ матриця, що нагадує словник.

$$E = \begin{pmatrix} 0{,}78 & 0{,}14 & 0{,}31 & 0{,}20 \\ 0{,}12 & 0{,}93 & 0{,}45 & 0{,}32 \\ 0{,}21 & 0{,}18 & 0{,}67 & 0{,}89 \\ 0{,}38 & 0{,}19 & 0{,}24 & 0{,}30 \\ 0{,}15 & 0{,}48 & 0{,}29 & 0{,}25 \end{pmatrix} \tag{2.1}$$

Тут представлено кожне слово у лексиціза одним рядком. Більше того, E – бажаний вихід кожного завдання, яке спрямоване на створення словауявлення. Вбудовування для третього слова, яке $w_3 = “fox”$ на ілюстрації, зберігається в третьому ряду E.

$$v_3 = (0,0,1,0,0) \times E = (0{,}21, 0{,}18, 0{,}67, 0{,}89) \quad (2.2)$$

Можна отримати, помноживши однозначне кодування слова на E. Для подальшого використання. Вкладення, одна проблема може бути в обробці слів, які не містяться в E. Викликані з лексики слова. Тим часом існує багато різних методів створити вбудовування слів. Word2Vec, GloVe та fastText – найпопулярніші підходи.

Модель word2vec була введена Міколовим та ін. у 2013 р. Ідея полягає в тому, щоб перетворити слово в безперервний вектор, який також представляє локальний контекст слова. Word2Vec робить ці означення більш відмітними, включаючи слова попередники і наступники.

Початковий документ word2vec пропонує два різні підходи до вивчання словосполучення з текстових корпусів із лексикою N. Перший підхід – це Continuous Bag-of-Words, яке передбачає слово, виходячи з його контексту. Другий підхід називається Skip-gram і прогнозує контекст слова. Обидві спроби мінімізують обчислювальні складності і ґрунтуються на нейронній мережі прямого поширення.

Для обчислення векторного зображення слова нейронна мережа CBOW приймає контекст терміну з лівого w_{t-2}, w_{t-1} і правого w_{t-2}, w_{t+2} слова w_t як вхідні дані. Це вікно слова може бути скоригована для збільшення або зменшення розміру локального контексту, пов'язаного зі слово будуванням t . На практиці контекстне вікно 5 для CBOW та 10 для Skip-gram [13]. CBOW приймає кодування кожного введеного слова і передає їх до мережі, щоб передбачити одноразове кодування цільового слова w_t. У неронній мережі прямого поширення моделює місце у вихідному векторі, що можна розглядати як завдання класифікації.

Мережа, яка використовується у Skip-gram, міняє вхід та вихід. Цільове слово тепер вхідне повідомлення і нейронна мережа прямого поширення передбачає контекст слова. Пропуск грам призводить до кращих результатів для невеликих корпорацій, в той час як CBOW є більш ефективним та рекомендованим для великих наборів текстів. Для тренування векторів слів; обидва підходи використовують багатошаровий перцептор.

Потрібні нам дані будуть виходити так. Кожна дужка позначає одиничне контекстне вікно. Синє поле позначає вхідний one-hot вектор, червоне поле – вихідний one-hot вектор. З одного контекстного вікна виходять два елементи даних. Розмір вікна зазвичай визначається користувачем. Чим більше розмір контекстного вікна, тим краще наша модель, але це впливає на час виконання алгоритму. Не треба плутати цільове слово з цільовими даними, це абсолютно різні речі.

Можна помітити, що всі сучасні програми NLP грунтуються на алгоритмах word2vec. Також можна поліпшити існуючі моделі векторними уявленнями слів. Моделі дозволяють відобразити семантично подібні слова в близькі один одному вектора в деякій моделі, в той час як далекі за змістом слова будуть виглядати по-різному. Це бажане властивість моделі, яке приведе до кращого результату.

3 РЕЗУЛЬТАТИ ТЕОРЕТИЧНИХ ДОСЛІДЖЕНЬ

Природною мовою називається будь-яку мову, який використовує людина. Мова може приймати різні форми, наприклад рукописний текст. Метою обробки природної мови є побудова такої системи, яка б «розуміла» людську мову і могла б керувати отриманими даними.

3.1 Розвиток завдання NLP

Історично склалося так, що обробка мови і мови розглядалась дуже по-різному в інформатиці, електротехніці, лінгвістиці. Через це розмаїття обробка мови і мови охоплює ряд незбіжних, але перекриваються областей різних дисциплін: комп’ютерна лінгвістика в лінгвістиці, обробка природної мови в інформатиці, розпізнавання мови в електротехніці, комп'ютерна психолінгвістика психології.

Вважається, що історія обробки природної мови бере початок другій половині двадцятого століття. У 1950 році Алан Тьюринг опублікував статтю під назвою «Обчислювальна техніка та інтелект», в якої в якості критерію інтелекту був запропонований так званий тест Тьюринга.

У 1960-х роках була розроблена досить успішна система оброблення природної мови: ELIZA. Це система написана Джозефом Вайзенбаумом між 1964 і 1966 роками і є симуляцією психотерапевта. ELIZA моделювала розмову, використовуючи зіставлення шаблонів і методологію заміни, яка давала користувачу ілюзію розуміння з боку програми. Наприклад при отриманні тексту «У мене болить голова » ELIZA переформулювати його в «Чому ви говорите, що у вас болить голова?". У 1970-ті роки багато програмісти почали писати «концептуальні онтології », які структурували реальну інформацію в зрозумілі для

комп'ютера дані. До 1980-х років більшість систем обробки природної мови ґрунтувалися на складних наборах рукописних правилах. Однак, починаючи з кінця 1980-х років, відбулася революція в обробці природної мови з впровадженням алгоритмів машинного навчання для обробки мови . Починаючи з 1990-х NLP стрімко розвиватися [14]. Величезна кількість текстів, які наводнили Інтернет, стимулювали роботу над завданнями з управління цими даними, зокрема шляхом добування інформації та автоматичного анотування.

У двадцять першому столітті, завдяки збільшенню обчислювальної потужності, низької вартості і доступності комп'ютерної пам'яті, активно розвиваються розробки в області NLP. Сучасні розробки породили безліч нових і складних дослідних тем. Починаючи від машинного читання тексту, в якому комп'ютери можуть «читати» і «розуміти» текст, до розпізнавання мови. Комп'ютери по всьому світу стає все більше і більше просунутими, наприклад: Watson – це комп'ютерна система відповідей на питання, здатна відповідати на питання, задані природною мовою, розроблена в IBM в 2006 році. Watson вислуховує питання, «розуміє» його, використовуючи сучасні методи NLP, проводить пошук по масивної базі даних, і знаходить найбільш правильну відповідь.

3.2 Проблематика обробки природної мови

Природна мова є нашим основним засобом комунікації і може бути охарактеризований як набір символів, які організуємо в структурній формі в висловлювання. дані висловлювання передають певний сенс. Структура висловлювань називається його синтаксисом, а сенс – семантикою. Області комп'ютерної лінгвістики NLP пропонують велику різноманітність уявлень і формалізмові для опису мови. Фактично, питання про те, як концептуалізувати

мову, тісно пов'язаний з дослідженнями в області людського пізнання, математичної логіки і філософії.

Дві проблеми ускладнюють обробку природних мов - рівень невизначеності, який існує в природних мовах та складність семантичної інформації, яка міститься навіть в простих реченнях [15].

Зазвичай мовні процесори мають справу з великою кількістю слів, багато з яких мають альтернативне значення. Складність, в тому числі, полягає в нерівномірності мови і в різних видах неоднозначності, які виникають в тексті.

Розглянемо деякі приклади.

У висловлюванні «Мама мила скло» існує проблема визначення частини мови слів. Слово «мила» може бути використано як в якості дієслова, який позначає дію «мити», так і в якості іменника, яке позначає предмет «мило». Будь-який механізм розбору тексту повинен бути здатний досліджувати різні синтаксичні конструкції фраз і мати можливість відстежувати і перебудовувати їх у міру необхідності.

Розглянемо два висловлювання: «Дитина штовхнув м'яч» і «Дитина штовхнула стіну». У першому висловлюванні сенс полягає в тому, що «дитина» виконаоа дію «штовхнула», яке привело в рух предмет «м'яч». У другому висловлюванні смислове навантаження інше – «дитина» виконана дію «штовхнула», тобто привів в рух «ногу», але слово «нога» не вказано явно в реченні. В даному випадку має місце неоднозначність. У вирішенні цієї неоднозначності може допомогти знання про те, що м’яч є рухомими об'єктами і часто переміщаються за допомогою ноги в якості інструменту, а стіни – статичними об'єктами.

Складнощі в обробці мови також пов'язані з особливостями і раз особистими лінгвістичними явищами мов. Розглянемо проблеми, котрі мають місце при аналізі україномовних текстів.

Вільний порядок слів. При аналізі висловлювань «Сашко кохає Марію» і «Марія кохає Сашка» результат буде різний. Обидва ланцюжка складаються з одних і тих же лексем, але в першому випадку, акцент робиться на те, що саме

«Сашко» «кохає Марію». У другому випадку, наголос йде на те, що саме «Марія» здійснює дію «кохати» по відношенню до «Сашка».

Розкриття анафор. Розглянемо два висловлювання «Сашко приніс Марії тарілку пастили, тому що вона його просила» і «Сашко приніс Марії тарілку пастили, тому що вона смачна» – дані висловлюють схожу структуру. У лівій частині висловлювань згадуються дві особи – «Сашко» і «Марія». У першому випадку займенник «вона» в правій частині вказує на «Марію», але в другому випадку займенник «вона» ставиться до слова «пастила». Процес розкриття анафор в українській мові є трудомістким і складним.·

Пунктуація. Залежно від розташування ком в реченні може змінюватися сенс висловлювання. Розглянемо крилатий вислів «Стратити не можна помилувати». Сенс фраз «стратити не можна, помилувати» і «стратити, не можна помилувати» протилежний.

Омоніми – це однакові за написанням, але різні за значенням слова. Як приклад можна привести «Маша з косою працювала в полі косою». В даному випадку слово «коса» розглядається як варіант зачіски «коса», і як значення інструменту «коса».

Представлені приклади показують, які виникають складності при обробці природної мови – тексту. Системи NLP змушені вирішувати проблеми, аналогічні тим, що були проілюстровані вище.

3.3 Загальні етапи обробки тексту

Мова може бути визначений як набір символів. символи об'єднуються і використовуються для передачі інформації або трансляції інформації. В процесі обробки природної мови виділяють чотири основних етапи. У реальних системах ці етапи рідко відбуваються як окремі, послідовні процеси. Деяких системах

частина етапів відсутня, об'єднується або вводяться додаткові кроки обробки тексту.

Токенізація і сегментація – розбиття тексту на токени слова і речення. Перехід від символів до речень і до слів.

Морфологічний аналіз – здійснює аналіз словоформ і позов їх лексем. Перехід до основ слова.

Синтаксичний аналіз – аналіз структурних відносин і зв'язків між словами.

Семантичний і прагматичний аналіз – аналіз смислової складової тексту.

У лінгвістичному аналізі текстів на природній мові необхідно чітко визначати, що є слово і речення. Розподіл цих одиниць породжує різні завдання, але ці завдання є нетривіальними, враховуючи різноманітність людських мов і систем писемності. Слова і речення, виявлені на даному етапі, є основними одиницями, прийнятими для подальшої обробки тексту.

Токенізація – це процес розбиття послідовності символів в тексті шляхом визначення меж слова, де закінчується одне слово, і починається інше [16]. У комп'ютерній лінгвістиці ідентифіковані таким чином слова називаються токенами. Даний етап важливий і подавати складність в письмових мовах, де в системі письма немає явних кордонів слів.

Токенізація слів може здатися простою, наприклад, в українській мові, де слова поділяються спеціальним символом – пропуском. Але в таких мовах як китайський, японський немає такої системи кордонів слів. До того ж, поділ слів тільки пропуском недостатньо. Розглянемо речення: «Маша мила посуд, вона втомилася.» Якщо розглядати слова, як то ,що знаходиться між кордонів – прогалин, то тоді при обробці даного прикладу виникнуть слова «посуд,», «втомилася». Ці помилки можна усунути, розглядаючи пунктуаційні знаки як доповнення до пробілу. Проте це може привести до нових помилок, так як неправильно будуть ідентифіковані наступні слова: «тощо», «spbu.ru», «36.6», «21/09/93» і так далі. Алгоритми токенізації також здатні розділяти багато немов вираження, такі як «Йошкар-Ола», для чого потрібно словник багатослівних

виразів. Це робить токенізацію тісно пов'язаної з задачею виявлення імен, дат і організацій, тобто з завданням розпізнавання іменованих сутностей.

Сегментація – це процес розбиття тексту на пропозиції. Поділ тексту на пропозиції зазвичай засноване на пунктуації. Це пов'язано з тим, що деякі види розділових знаків, точки, знак запитання, знаки оклику, як правило, позначають кордони пропозиції. Знаки запитання й оклику є однозначними маркерами кордонів пропозицій, а символ «.» не завжди означає кордон пропозиції. Тому поділ тексту на пропозиції і слова, як правило, розглядаються спільно.

Морфема – найменша значуща мовна одиниця. Морфеми діються на два основних типи: корінь і афікс. Це відповідно основна значуща і допоміжна частини слова. Афікси розділяються на два типи: префікси і постфікси, які розташовуються відповідно до і після кореня. В українській мові префікси – це приставки, що ставляться перед коренем слова, а постфікси – це суфікси і закінчення.

Морфологічний аналіз – це процес зіставлення словоформам їх лексем, а також перехід до основних значимих частина слова. Морфологічний аналіз в першу чергу займається визначенням того, як була змінена «базова» форма слова. Зміна зазвичай відбувається шляхом додавання префіксів і постфіксів. Наприклад, словоформа «нещасливому» має лексему «нещасливий», приставку не-, корінь счаст-, суфікс -лів і закінчення -ому.

Характер морфологічного аналізу в значній мірі залежить від мови з якої робиться аналіз. У деяких мовах окремі слова містять всю інформацію про час, рід і число так далі. В інших мовах ця інформація може бути передана через кілька слів у реченні. Наприклад, в англомовному пропозиції «He will have wrote this essay by Monday.» складна інформація про час, що передається через допоміжні дієслова «will» і «have». В українській мові пропозиції «Він напише твір до понеділка.» інформація про час передається за допомогою афікса дієслова «напише».

Морфологічний розбір дуже важливий. Він грає вирішальну роль в задачі інформаційного пошуку в морфологічно складних мовах, таких як українська або німецька. В українській мові існує безліч словоформ однієї і тієї ж лексеми. Як приклад схиляння слова «стіл» по відмінку і числу. Таким чином, наприклад,

завдання пошуку інформації про слово «стіл» в тексті повинна враховують всілякі варіанти зміни цього слова.

Морфологічний аналіз – це процес знаходження складових морфем слова. На даному етапі здійснюється POS-тегування, тобто для кожного слова в тексті визначається частина мови і набір морфологічних характеристик.

Мета синтаксичного аналізу складається з двох частин: перевірити, що рядок слів сформована коректна і створити структуру, показує синтаксичні відносини між різними словами. Синтаксичний аналізатор робить це, використовуючи словник визначень слів(Лексикон) і набір синтаксичних правил [17]. Проста лексика містить тільки синтаксичну категорію кожного слова, проста граматика описує правила, які вказують тільки як синтаксичні категорії можуть бути об'єднані для формування фраз різних типів. Не всі системи NLP вимагають повного розбору пропозицій, тобто проведення повного синтаксичного аналізу.

Семантичний та прагматичний аналіз. Кінцева мета, як для людини, так і для системи обробки природної мови, полягає в тому, щоб зрозуміти висловлювання. «Розуміння» висловлювання – це складний процес, який залежить від результатів попередніх етапів системи NLP, а також від лексичної інформації, контексту і здорового глузду; і призводить до того, що називається семантичною інтерпретацією в контексті висловлювання.

Проектування семантичного інтерпретатора передбачає рішення тих же проблем, з якими доводиться стикатися при побудові синтаксичного аналізатора, зокрема, проблеми семантичної двозначності, як розпізнати передбачувану семантичну інтерпретацію висловлювання пропозиції в конкретному контексті, серед множини можливих інтерпретацій цієї пропозиції.

Завдання розробки семантичного інтерпретатора складне, оскільки між дослідниками мало згоди щодо того, який саме повинна бути кінцева інтерпретація висловлювання. Практичні системи NLP, як правило, використовують семантичні уявлення, що призначені для конкретної предметної області.

4 МЕТОДИ МАШИННОГО НАВЧАННЯ

Дослідники в області машинного навчання розглядають питання про те, як зробити машини здатними «вчитися», тобто щоб машини мали здатність змінювати свою поведінку таким чином, щоб в майбутньому виконувати завдання краще. Зокрема, машинне навчання – це метод створення комп'ютерних програм шляхом аналізу даних. Деякі системи ML засновані на взаємодії людей і машин, в той час як інші намагаються усунути потребу в людині при аналізі даних. Машинне навчання є результатом перетину інформатики та статистики.

Методи машинного навчання широко застосовуються в інтелектуальному аналізі даних, зокрема, для виявлення прихованих закономірностей в зростаючих обсягах даних в Інтернеті. У NLP і TM методи ML застосовуються для обробки масивних обсягів даних, що сприяє розвитку різних методів і прийомів для вирішення проблем, які виникають при обробці природної мови.

Відправною точкою для вирішення більшості завдань NLP і TM є пошук і створення структур в неструктурованих даних, що є основною метою методів класифікації і кластеризації.

4.1 Завдання класифікації

Завдання класифікації передбачає визначення, до якого класу або класів із заздалегідь заданого набору класів відноситься об'єкт, що аналізується. Цей процес аналогічний тому, як книгам в бібліотеці присвоюються різні категорії.

Одним з підходів до реалізації даного процесу є класифікація з використанням правил, які найчастіше формуються людиною. Такий набір правил фіксує певну комбінацію ключових слів, яка вказує на клас. Оскільки підтримка створених вручну правил є трудомістким процесом, існує ще один підхід до

класифікації тексту, а саме класифікація тексту на основі методів машинного навчання [18]. У ML критерії прийняття рішення щодо класифікації документа визначаються автоматично на основі даних, отриманих на етапі навчання. Навчальні дані повинні служити прикладом документів кожного класу. Дані документи повинні бути поміченими. Маркування даних є більш простим завданням, ніж написання правил. Цей тип ML називається навчанням з учителем, томущо для управління процесом навчання необхідний експерт.

Розглянемо задачу класифікації документа. Нехай $d \in D$, де d – деякий документ (від англ. document), D – безліч документів. Нехай $C = \{c\,1\,, c\,2\,, \ldots, c\,\}$, де $ci\,, i = 1, \ldots, n$ – певний клас. Безліч C – задається експертом вручну. також існуєнавчальна множина T, в якому містяться раз-мічені документи, тобто $< d, c > \in D \times C$.

Завдання класифікації полягає в пошуку функції класифікації γ ,за допомогою навчальної множини T і алгоритму навчання, наступного виду:

$$\gamma\colon D \rightarrow C \tag{4.1}$$

4.2 Завдання кластеризації

На відміну від класифікації, кластеризація є найбільш поширенимою формою неконтрольованого навчання. Немає людини-експерта, який повинен визначити, які об'єкти до яких класів відносяться. Алгоритми кластеризації групують набір об'єктів в підмножину кластерів. Мета алгоритмів полягає в тому, щоб створити кластери. Внутрішні елементи кожного кластера повинні бути схожі між собою і при цьому повинні відрізнятися від елементів інших кластерів.

Розглянемо кластеризацію об'єктів – точок з урахуванням їхнього економічного становища відносно один одного, які зображені на рисунку 4.1.

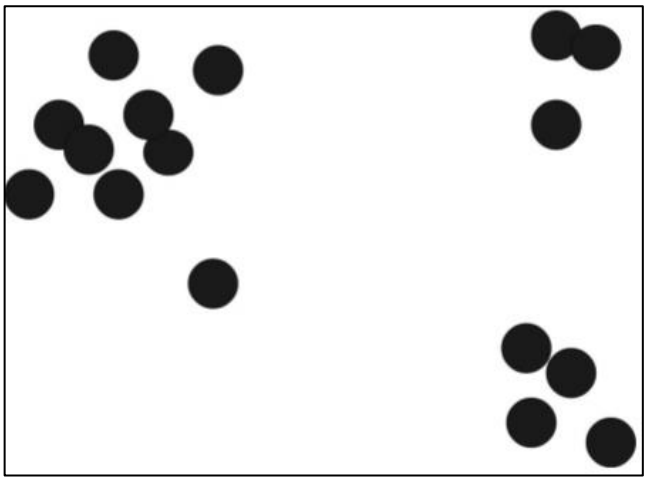

Рисунок 4.1 – Вхідні дані.

В результаті застосування алгоритму кластеризації ми можемо відокремити три кластера A , B , C, зображені на рисунку 4.2, всередині яких знаходяться схожі між собою елементи.

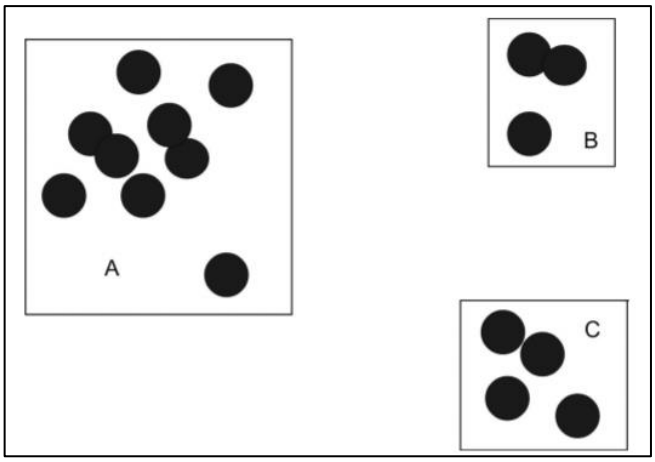

Рисунок 4.2 – Кластери A, B, C.

Нехай $D = \{d\,1\,, d\,2\,, \ldots, d\,\}$ – безліч документів. Нехай К $= \{1,2,\ldots, N\}$ безліч ідентифікаторів кластерів, а N – число, яке позначає загальна кількість кластерів (в загальному випадку заздалегідь невідомо). Нехай є функція $\rho(d1, d2)$, яка позначає ступінь схожості двох документів $d1, d2 \in D$. Завдання кластеризації полягає в тому, щоб знайти мінімально можливе N, при якому кожному елементу $d \in D$ будепоставлений у відповідність $k \in$ К, з урахуванням близькості метрики ρ елементів кластера.

4.3 ML в завданні вилучення інформації

Вилучення інформації визначається як ідентифікація визначених сутностей і відносин. З точки зору машинного навчання, це визначення відповідає класифікації фрагментів тексту згідно заданого класу або мітки. Таким чином, існує можливість для представлення завдання добування інформації як завдання машинного навчання. Методи машинного навчання для отримання інформації зазвичай використовують моделі з навчанням. Параметри моделей оцінюються відповідно до наведених анотованих навчальних даних для прогнозування набору міток або сутностей.

Прикладом може служити завдання розпізнавання іменованих сутностей. Розглянемо вхідний текстовий файл, в якому міститься розмічений текст. Слова, які позначають імена сутностей, позначені «1». Всі інші слова позначені «0». Таким чином, завдання вилучення імен з тексту може розглядатися, як процес класифікації віднесення кожного слова до одного з двох класів «1» і «0». У зв'язку з цим, алгоритм класифікації можна розглядати як алгоритм, що використовується для того, щоб визначити, які підстроки тексту потрібно витягнути, а які ні.

Простим підходом до класифікації є використання байєсівського класифікатора [19]. У машинному навчанні байєсівські класифікатори представляють собою сімейство простих верогідностних класифікаторів, заснованих на застосуванні теореми Байеса із строгими припущеннями щодо незалежності між ознаками.

Незважаючи на свою простоту, вони широко використовуються в ML і NLP з великим успіхом. Байєсовські класифікатори припускають, що вплив значення змінної на даний клас не залежить від значень інших змінних. Це припущення називається умовною незалежністю класу. Припущення зроблено для спрощення обчислень і в цьому сенсі вважається наївним. Дослідження показали високу точність і швидкість роботи цих методів при застосуванні до великих наборів даних.

Маркування послідовності використовується в багатьох задачах обробки природної мови. Слова тексту розглядаються як спостерігання, які мають маркування – набір характеристик. В цьому випадку змінюються моделі послідовностей: прихована Марківська модель, Марківська модель з максимальною ентропією, метод умовних випадкових полів.

Завдання класифікації документів полягає в тому, щоб спів віднести документ з набору до груп або групи документів, які є схожими між собою. Суть груп документів може бути різна, вони можуть являти собою конкретні теми, певні події чи категорії. У разі машинного навчання, для вирішення цього завдання необхідно виконати навчання з вчителем. Для навчання з вчителем системі необхідно надати навчальну вибірку, що складається з певного набору текстових документів, а також набору класів, до яких ці текстові документи належать.

У разі кластеризації документів, побудована система повинна самостійно встановити безліч різних кластерів, до яких можуть відноситися текстові документи. Це завдання в термінах машинного навчання іменується навчанням без вчителя.

Завдання вилучення інформації полягає в ідентифікації зумовлених об'єктів з тексту, наприклад сутностей або відносин. З точки зору області ML це відповідає завданню класифікації. Таким чином, можна розглядати вилучення інформації як завдання ML.

5 РЕАЛІЗАЦІЯ АВТОМАТИЗОВАНОЇ СИСТЕМИ

У цьому розділі представлена реалізації автоматизованої системи визначення події міжнародного значення на основі аналізу стрічок на різних мовах.

Сформулюємо умови, які накладаються на систему визначення подій. В якості вихідних даних виступають новинні повідомлення особистих засобів масової інформації. В системі використовуються джерела інформації двох типів: новинна стрічка користувача і альтернативна стрічка. Новинна стрічка користувача будується на основі популярних ЗМІ, які поширюються вільно в Інтернеті. Як альтернативну стрічки розглядаються новинні стрічки на англійській, французькій та німецькій мовами.

Найчастіше новинні статті складаються з наступних блоків: заголовок, новинне повідомлення, дата, додаткові позначки. Функція новинного заголовка полягає в тому, щоб залучити увагу читача [20]. Найчастіше, в заголовку коротко і ємко представлена ключова інформація новинний статті. В рамках даної роботи розподіл події відбувається в межах одного речення, на основі аналізу заголовка новинного повідомлення мовою.

Під «подією» будемо розуміти набір ключових слів, які відповідають на питання «хто і що зробив?». Уявімо подія у вигляді трьох ключових слів: суб'єкт (Subject), предикат (Predicate), об'єкт (Object). Під предикатом будемо розуміти дію, котрої направлено суб'єктом по відношенню до об'єкта. Відзначимо, що в мові суб'єкту в реченні відповідає підмет, предикату – присудок, об'єкту – доповнення. Будемо вважати, що подія ідентифіковано, якщо в трійці «Subject, Predicate, Object» знайдений хоча б предикат. Подія вважається міжнародною, якщо в найденій статті про дану подію в альтернативній новинній стрічці. При пошуку новин викриються ключові слова, які перекладені мовами альтернативної новинної стрічки.

5.1 Архітектура системи

Схематично архітектура системи представлена на рисунку 5.1.

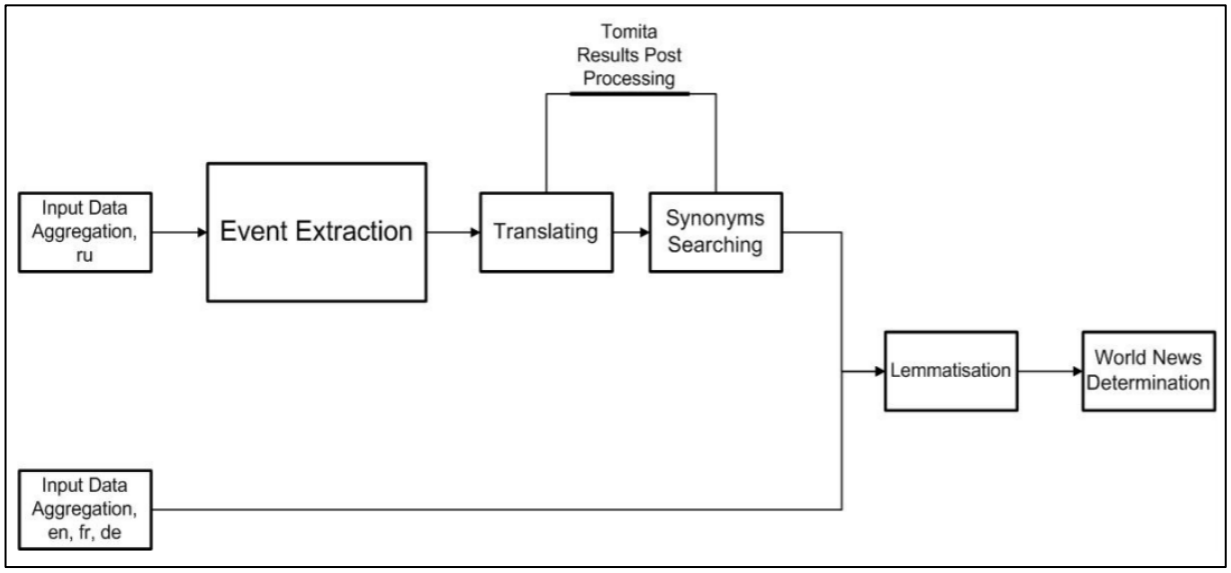

Рисунок 5.1 – Архітектура системи.

Система складається з п'яти модулів:

- Input Data Aggregation – у цьому модулі відбувається збір вихідних даних;
- Event Extraction – ключовий модуль системи, в якому проходить обробка тексту: токенізація, сегментація і морфологічний аналіз;
- Tomita Results Post Processing – модуль піст обробки результатів, отриманих на попередньому етапі, модуль складається з двох компонентів: переклад і пошук синонімів;
- Lemmatisation – модуль лематизації, тобто приведення слів до їх лем;
- World News Determination – модуль визначення міжнародної новини та надання альтернативних новин.

5.2 Модулі системи

Як стрічці новин користувача розглядаються наступні джерела інформації – UKR.NET, Gazeta.ua, Уніан. Як альтернативні: англомовні – The Guardian, BBC News, The Washington Post, The New York Times, USA Today, TheIndependent; французьомовні – Le Figaro, Le Monde, Libération; німецькомовні – Deutsche Welle, Frankfurter Allgemeine Zeitung, Die Welt.

Збір інформації здійснюється на основі RSS-каналів новинних порталів. RSS – сімейства XML-форматів, які спеціалізуються на описі стрічки новин. Ці канали можуть, наприклад, дозволити користувачеві стежити за безліч різних веб-сайтів в одному зібранні новин. Стандартні формати файлу XML забезпечує сумісність з багатьма програмами.

За рахунок того, що RSS – це загальноприйнятий і стандартизований формат передачі інформації, усі отримані на виході XML файли мають ідентичну структуру, незалежно від джерела поширення. Для всіх XML файлів можна застосовувати однаковий механізм десеріалізації. Десеріалізація – процес відновлення вихідної структури даних за допомогою послідовності бітів. У даній роботі використовувався пакет System.ServiceModel.Syndication, призначений для обробки XML, використовуваних в RSS розсилках.

В результаті десеріалізації ми отримуємо список об'єктів, кожен з яких представляє конкретну новину з довільної RSS стрічки.

Для реалізації модуля вилучення інформації використовується інструмент Томіта-парсер. Витяг даних здійснюється з допомогою шаблонів і з використанням словників ключових слів. Парсер вирішує необхідні завдання обробки тексту: токенізацію, сегментацію, морфологічний аналіз. В системі відсутній візуальне середовище розробки, управління здійснюється за допомогою командного рядка.

Подія – «Суб'єкт, Предикат, Об'єкт». Кожне з полів може містити в собі кілька слів. Для однієї події може бути знайдено кілька полів. Так як основна функція заголовка полягає втому, щоб привернути увагу, ключова інформація знаходиться на початку заголовка. Найчастіше новинний заголовок будується

наступним чином: «Суб'єкт, Предикат, Об'єкт, Решта слова» або «Об'єкт, Предикат, Суб'єкт, Решта слова». Ця особливість враховується при побудові модуля вилучення події.

На вхід програмі подається txt файл заголовків новин. Під предикатом будемо розуміти «дію». Під суб'єктом будемо розуміти «того, хто здійснює дію». Під словом «група» будемо мати на увазі ланцюжок поспіль слів, задоволених вказаною умові.

Під об'єктом будемо розуміти «того, по відношенню до кого здійснюють дію». Таким чином, витяг об'єктів з тексту аналогічно задачі вилучення суб'єкта, без прив'язки до називному відмінку. Для роботи парсера потрібні такі вихідні файли: config.proto конфігураційний файл, dic.gzt – кореневий словник, grammar.cxx – граматики, fact_types.proto – файл опису типів фактів, kw_types.proto файл опису типів ключових слів. Файл grammar.cxx містить в собі правила, написані на мові контекстно-вільних граматик.

Модуль складається з двох компонент: переклад і пошук синонімів. В результаті роботи попереднього модуля є список новинних заголовків і витягнуті з нього трійки «Суб'єкт, Предикат, Об'єкт», які можуть складатися з декількох полей, всередині яких знаходяться ключові слова. Для подальшого пошуку ключових слів в альтернативній стрічці їх необхідно перекласти.

Переклад ключових слів здійснюється за допомогою API Гугл перекладача. Клієнт відправляє http-запит на сервер, сервер повертає відповідь клієнту. HTTP протокол запиту та відповіді в моделі клієнт-сервер [21]. Для перекладу, в тілі HTTP запиту до API необхідно вказати текст, який треба перевести, мова з якої здійснюється переклад і мову призначення перекладу.

Для поліпшення ефективності пошуку ключових слів в альтернативній стрічці, а також для вирішення завдання, проводиться пошук синонімів. Модуль реалізується з використанням публічного API, що надається сервісом thesaurus.altervista. Даний сервіс дозволяє знаходити синоніми для багатьох мов, на основі відкритих словників OpenOffice.

Для пошуку ключових слів в альтернативних джерелах інформації необхідно привести їх до нормального вигляду, тобто провести лематизацию. Лематизація – це перехід від словоформи до лема слова. Крім ключових слів, проводиться лематизації новинних повідомлень з альтернативної стрічки.

Модуль лематизації реалізований на основі бібліотек OpenSource проекту LemmaGenerator. Даний сервіс виконаний за рахунок зберігання основних наборів лем для різних мов і подальшого аналізу необхідного тексту на предмет їх відповідності. У даній роботі використовувався поставляється разом з лематизатором набір списків лем.

В результаті роботи попередніх модулів на вхід модуля World News Determination подається два типи даних. Зазначені слова переведені намови альтернативної новинної стрічки і нормалізовані. Заголовки і змісту новин альтернативної стрічки. Зазначені дані нормалізовані.

Для визначення міжнародної події проводиться пошук ключових слів в заголовках і повідомленнях альтернативної стрічки. Здійснюється пошук одного з полів «Суб'єкт, Предикат, Об'єкт» і «Суб'єкт, Предикат», «Предикат, Об'єкт», «Суб'єкт, Об'єкт».

ВИСНОВОК

Ця магістерська робота представила сучасні методи в NLP і показала підхід до застосування цих методів. Поточні дослідження аналізу тексту на основі NLP зосереджені на вдосконаленій штучній нейронній мережі твори, які використовують значущі подання слів як вхідні дані. Ця робота охоплювала нейронну мережу прямого розповсюдження, рекурентну нейронну мережу, на основі уваги рекурентна нейронна мережа, ієрархічна модель мережі, які використовуються для аналізу текстів стосовно даної цільової змінної. Крім того, представлено різні типи слів вбудовування та використовували їх у підході, щоб показати свою ефективність на створеному користувачем тексти. Перетворення слів у змістовний та формати для мишиного читання, суттєво з'являється старе дослідження в NLP, що спирається на складні архітектури глибокого навчання на правлені рекурентні нейронні мережі. Забезпечивши теоретичні основи та огляд літератури.

Основну увагу було приділено класифікації тексту, оскільки це дало найкращі рішення для аналізу текстів щодо їх впливу. Більше того, дослідження показали, що аналіз текстів опису лише для прогнозують результат бізнесу, такий як ціна, в літературі досить рідко висвітлено.

У даній роботі продемонстровано можливість побудови автоматизованої системи визначення міжнародної події і реалізований інструмент надання кінцевому користувачеві інформації про рівень значущості події.

В ході роботи були розглянуті та вивчені різні засоби і підходи до інтелектуального аналізу тексту для коректного вилучення події з російськомовної новини. Також була представлена реалізація запропонованої системи з використанням Томіта-парсера реалізація повністю відповідає поставленим перед нею вимогам:

- збір вихідних даних;
- обробка новинних документів;

- визначення та витяг подій з новини;
- перевірка події, для визначення чи є подія міжнародним;
- надання користувачеві новини про ідентифікованому подію з альтернативних джерел.

Був проведений аналіз ефективності роботи побудованої системи: точність і повнота рівні 0.86 і 0.79 відповідно. З чого можна зробити висновок про те, що даний підхід має позитивні результати і вирішує поставлене завдання в достатній мірі для практичного застосування

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных: навч. посіб. – М.: НИУ ВШЭ, 2017. – 269 с.

2. Четвериков Г.Г., Дударь З.В., Вечирская І.Д., Дискретні структури: навч. посібник для студентів, які навчаються за напрямами "Комп'ютерні науки" і "Прикладна математика". – Харьків: ХНУРЄ, 2014. – 319 с.

3. Poibeau T., Saggion H., Piskorski J., Yangarber R. Multi-source, Multilingual Information Extraction and Summarization. Theory and Applications of Natural Language Processing. Springer, Berlin, Heidelberg, 2013. – 257 с.

4. Ted Briscoe. Introduction to Linguistics for Natural Language Processing. Tech. rep. 2013, pp. 1–37.

5. Морозов Н.А. Лінгвістичні спектри: засіб для відрізнення плагіатом від справжніх творів того чи іншого відомого автора. Стілеметріческій етюд. // Известия отд. російської мови і словесності Імп.Акад.наук, Т.XX, кн.4, 1915.

6. Марков А.А. Приклад статистичного дослідження над текстом "Євгенія Онєгіна", який ілюструє зв'язок випробувань в ланцюг. // Известия Імп.Акад.наук, серія VI, Т.X, N3, 1913 – 153 с.

7. Марков А. А. Про один застосуванні статистичного методу. // Известия Імп.Акад.наук, серія VI, Т.X, N4, 1916 – 239 с.

8. Мартиненко Г. Я. Основи стилеметрії. – Л.: Вид-во ЛДУ, 1988. – 176 с.

9. Хмельов Д.В. Розпізнавання автора тексту з використанням ланцюгів А.А. Маркова // Вести. МГУ. Сер. 9. Філологія. 2000. №2. С. 115-126.

10. Шитіков В. К., Мастіцкій С. Е. Класифікація, регресія і інші алгоритми Data Mining з використанням R. 2017. – 253 с.

11. Від Нестора до Фонвізіна. Нові методи визначення авторства. М.: Издат. група «Прогрес», 1994. – 293 с.

12. Паничев В.В., Солов'їв Н.А. Комп'ютерне моделювання: навч. посіб. – Оренбург: ГОУ ОГУ, 2008. – 130 с.

13. Девіс К.Х., Біддульф Р. та Балашек С. Автоматичне розпізнавання мовлення розмовних цифр, 1952. – С. 637- 642.

14. Т. Хасті, Р. Тібшірані, Дж. Фрідман. Елементи статистичного навчання. Збір даних, висновок та прогнозування. 2-е видання. – Спрингер, 2013. – 335 с

15. Gobinda G. Chowdhury. "Natural Language Processing". In: Annual Review of Information Science and Technology, 2003. – C. 51-89.

16. Junyoung Chung et al. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: 2014. pp 4–25

17. Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu. Deep Learning Based Text Classification: A Comprehensive Review // 2005.

18. Tom Young, Devamanyu Hazarika, Soujanya Poria, Erik Cambria. Recent Trends in Deep Learning Based Natural Language Processing // – 2018. pp 6–20

19. Cunningham H. Developing language processing components with GATE, Technical report / H. Cunningham, D. Maynard, K. Bontcheva et al.- University of Sheffield, U.K., 2005 [Електронний ресурс] Режим доступа: http://www.gate.ac.uk.

20. Delong G. An overview of the FRUMP system/ D.G.Bobrow, R.Kaplan, M.Kay et al. (1977) // In Lehnert W. and Ringle M. (Eds.), Strategies for Natural Language Processing, Lawrence Eribaum, Potomac, Maryland.- 1982.- p. 149-176.

21. Robert P Schumaker and Hsinchun Chen. "Textual Analysis of Stock MarketPrediction Using Breaking Financial News : The AZFinText System". In: 2006, pp. 1–29.