

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СУМСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ФАКУЛЬТЕТ ЕЛЕКТРОНІКИ ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА КОМП'ЮТЕРНИХ НАУК
СЕКЦІЯ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ПРОЕКТУВАННЯ

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему: «Програмний додаток визначення медичної дезінформації в
онлайн-спільнотах»

за спеціальністю 122 «Комп'ютерні науки»,
освітньо-професійна програма «Інформаційні технології проектування»

Виконавець роботи: студент групи ІТ.м-82 Вербицька Анастасія Анатоліївна

**Кваліфікаційну роботу
захищено на засіданні ЕК
з оцінкою**

«___» грудня 2019 р.

Науковий керівник

(підпис)

к.т.н., доц., Парфененко Ю. В.

Голова комісії

(підпис)

Шифрін Д.М.

Засвідчую, що у цій дипломній роботі немає
запозичень з праць інших авторів
без відповідних посилань.

Студент _____
(підпис)

Суми-2019

Сумський державний університет
Факультет електроніки та інформаційних технологій
Кафедра комп'ютерних наук
Секція інформаційних технологій проектування
Спеціальність 122 «Комп'ютерні науки»
Освітньо-професійна програма «Інформаційні технології проектування»

ЗАТВЕРДЖУЮ

Зав. секцією ІТП

_____ В. В. Шендрик
«__» _____ 2019 р.

ЗАВДАННЯ

на кваліфікаційну роботу магістра студентіві

Вербицька Анастасія Анатоліївна

(прізвище, ім'я, по батькові)

1 Тема проекту Програмний додаток визначення медичної дезінформації в
онлайн-спільнотах

затверджена наказом по університету від «19» листопада 2019 р. №2305-III

2 Термін здачі студентом закінченого проекту « 10 » грудня 2019 р.

3 Вхідні дані до проекту _____

4 Зміст розрахунково-пояснювальної записки (перелік питань, що їх належить розробити) _____

1 Аналіз прелметної області _____

2 Постановка задачі та методи дослідження _____

3 Моделювання програмного додатку _____

4 Реалізація програмного додатку _____

5 Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень) _____

Презентація (слайдів) _____

6. Консультанти випускної роботи із зазначенням розділів, що їх стосуються:

Розділ	Консультант	Підпис, дата	
		Завдання видав	Завдання прийняв

Дата видачі завдання _____.

Керівник _____
(підпис)

Завдання прийняв до виконання _____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Назва етапів випускної проекту	Термін виконання етапів проекту	Примітка
1	Аналіз предметної області	До 25.10.2019 р.	
2	Постановка задачі та методи дослідження	До 25.10.2019 р.	
3	Моделювання програмного додатку	До 19.11.2019	
4	Реалізація програмного додатку	До 10.12.2019 р.	
5	Оформлення пояснювальної записки до дипломного проекту	До 10.12.2019 р.	
6	Підготовка презентації	До 16.12.2019 р.	
7	Захист дипломного проекту	19.12.2019 р.	

Магістрант _____

Вербицька А.А.

Керівник роботи _____

к.т.н., доц. Парфененко Ю.В.

РЕФЕРАТ

Тема кваліфікаційної роботи магістра «Програмний додаток визначення медичної дезінформації в онлайн-спільнотах».

Пояснювальна записка складається зі вступу, 4 розділів, висновків, списку використаних джерел із 30 найменувань, додатків. Загальний обсяг роботи – 69 сторінок, у тому числі 48 сторінок основного тексту, 3 сторінки списку використаних джерел, 18 сторінок додатків.

В роботі проведено аналіз предметної області, визначено мету та основні задачі проекту та засоби реалізації. Проведено огляд інтелектуальних методів обробки природної мови та Python-бібліотек. У роботі виконано моделювання роботи програмного додатку за допомогою структурно-функціонального аналізу, побудовано та навчено нейронну мережу для класифікації тверджень та розроблено програмний додаток.

Результатом проведеної роботи є розроблений програмний додаток визначення медичної дезінформації в онлайн-спільнотах

Практичне значення розробленого додатка полягає у його використанні для перевірки публікацій та виявлення неправдивих даних при їх модерації.

Ключові слова: дезінформація, неправдиві новини, програмний додаток, нейронна мережа, згорткова нейронна мережа, класифікація тексту.

ЗМІСТ

Вступ.....	6
1 Аналіз предметної області.....	8
1.1 Антисоціальна поведінка онлайн та поняття фейкових новин.....	8
1.2 Визначення медичної дезінформації онлайн	11
1.3 Інтелектуальні методи обробки природної мови	13
2 Постановка задачі та методи дослідження	18
2.1 Мета та задачі.....	18
2.2 Огляд python-фреймворків для обробки природної мови	19
2.3 Засоби реалізації програного продукту	25
2.3.1 Формування набору даних для класифікації тверджень.....	26
2.4 Модель нейронної мержі для класифікації тверджень.....	28
3 Моделювання програмного додатку визначення медичної дезінформації в онлайн-спільнотах	33
3.1 Структурно-функціональне моделювання процесу визначення медичної дезінформації в онлайн-спільнотах.....	33
3.2 Побудова діаграми варіантів використання	37
3.3 Побудова діаграми діяльності	39
3.4 Логічне проектування моделі бази даних	41
4 Реалізація програмного додатку визначення медичної дезінформації.....	43
4.1 Архітектура роботи програмного додатку.....	43
4.2 Результати роботи.....	44
Висновки	50
Список літератури	51

ВСТУП

Швидкий розвиток мережі Інтернет привів до стрімкого зростання кількості різноманітних форумів, блогів, де користувачі спілкуються на різні теми, дають один одному поради. Наприклад, популярними є форуми, на котрих йдуть обговорення різних питань, що стосуються лікування та профілактики захворювань. Адже деяка частка людей все ж займається самолікуванням і замість того, щоб звернутися до спеціаліста, хоче швидко знайти відповіді на питання в Інтернеті.

Звичайно це може здатися швидким рішенням проблем, але потрібно завжди пам'ятати, що публікації на такого виду форумах проходять модерацію, яка зазвичай полягає у перевірці на наявність нецензурних слів та виявлення спаму. А проблема перевірки контенту, у тому числі на медичну тематику, залишається не вирішеною. Адже в більшості випадків модератори не мають прямого відношення до медичної сфери, і їм важко розпізнати чи правдива інформація в тій чи іншій статті, яку вони перевіряють, чи ні.

Тому, через велику кількість неправдивих новин у соціальних та медіа ЗМІ стала актуальною задача розроблення програмного додатка виявлення проявів медичної дезінформації в Інтернеті.

Об'єктом дослідження є процес визначення медичної дезінформації в онлайн-спільнотах.

Предмет дослідження – моделі і методи автоматизованого визначення медичної дезінформації онлайн-спільнотах.

Метою кваліфікаційної роботи магістра є розробка програмного додатка, що буде виявляти неправдиві новини та дезінформацію на медичну тематику в онлайн-спільнотах.

Для досягнення поставленої мети потрібно виконати ряд поставлених задач:

- дослідити предметну область, визначити поняття антисоціальної поведінки та неправдивих новин в онлайн-спільнотах;
- провести огляд інтелектуальних методів для обробки природної мови та обрати той, за яким буде проводитися розробка програмного додатка.
- розробити модель визначення антисоціальної поведінки;
- провести моделювання програмного додатка визначення медичної дезінформації в онлайн-спільнотах та виконати планування робіт з його розробки;
- розробити програмний додаток визначення медичної дезінформації в онлайн-спільнотах.

Практичне значення розробленого додатка полягає у його використанні для перевірки публікацій та виявлення неправдивих даних при їх модерації.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Антисоціальна поведінка онлайн та поняття фейкових новин

Антисоціальна поведінка – це поведінка, котра суперечить правовим, морально-етичним та культурним нормам. Основною ознакою даної поведінки є вчинення дій, що порушують принципи етики та моралі, є безвідповідальними та ігнорують закони та права інших людей [1].

Одним із проявів такої поведінки є антисоціальна поведінка в Інтернеті. На даний час веб-ресурси надають користувачам змогу дізнаватися безліч інформації, спілкуватися між собою та висловлювати свою власну думку. Але не завжди користувачі Інтернету поводять себе згідно з правовими та етичними нормами. За статистикою, 40 % користувачів Інтернету стикаються з неповагою інших антисоціальних користувачів [2].

Антисоціальна поведінка поділяється на два типи: створення та розповсюдження неправдивої інформації та дії антисоціальних користувачів.

До першого типу належать:

- неправдиві новини;
- чутки;
- необ'єктивні новини;
- неправдиві відгуки про товари та послуги;
- дезінформація.

Другий тип включає в себе:

- образливі висловлювання;
- тролінг;
- кібер-булінг;
- фейкові акаунти;
- спам;

- насильство з боку онлайн-спільноти.

На сьогоднішній день, в епоху неймовірної кількості інформації, неправдиві новини стали невід'ємною частиною нашого життя. Неправдиві новини – це частково або повністю неправдива інформація, що розповсюджується в соціальних медіа та ЗМІ, з метою введення в оману [3].

Розповсюдження на просторах інтернету так званих неправдивих новин на даний момент розглядається як одна з найбільших загроз демократії, журналістиці та свободі вираження поглядів. Адже за допомогою фейкових новин можна маніпулювати думкою людей, спонукати їх до певних не завжди правильних дій і навіть розв'язати війну. Основними джерелами розповсюдження неправдивої інформації є соціальні мережі (Facebook, Twitter, Instagram), Youtube, а також різні тематичні форуми, блоги, де спілкується велика кількість людей.

Дезінформація може зустрічатися на різну тематику. Не виключенням є і медична галузь. Лікування засобами, які радять користувачі форумів, а не призначають лікарі, є потенційно небезпечним.

Дослідники виділяють декілька типів неправдивих новин [4]

- Візуальний тип. В основі неправдивих новин даного типу використовуються зображення або відео з використанням фотошопу.
- Тип на основі користувача. Неправдиві новини даного типу орієнтовані на певну аудиторію певної вікової групи, статі чи культури.
- Неправдиві пости. Неправдиві новини, що зазвичай розміщуються на платформах соціальних медіа.
- Мережеві фейки. Мережеві новини, орієнтовані на членів певної організації, які так чи інакше пов'язані між собою.
- Неправдиві новини на основі знань. Даний тип новин містить наукове або розумне пояснення невирішеним питанням. Такі новини поширюють неправдиву інформацію, наприклад, помилкова стаття про те, яквилікувати астму.

- На основі стилю. неправдиві новини пишуть більшість людей, що не є журналістами, але при цьому стиль написання може бути різним.
- На основі позиції. Позиційно-орієнтований тип фокусується на тому, як висловлювання подаються в статті. Правдиві новини пишуться таким чином, щоб надати достатньо інформації про предмет, і читачі повинні розуміти зміст статті. Статті з неправдивими новинами ж написані для того, щоб дати мало інформації про предмет і зробити багато фальшивих заяв.

Визначити, чи є інформація фейком можна також, аналізуючи поведінку користувача. Наприклад, коли на форумі одна і та ж сама людина залишає багато однотипних повідомлень і проявляє велику активність, це одразу повинно викликати підозру. Також, проаналізувавши сам текст повідомлення чи статті, можна розпізнати неправдиву інформацію.

Існує декілька способів визначення дезінформації в онлайн-спільнотах. Перший спосіб полягає у ручній перевірці повідомлень, у тому числі із залученням профільних спеціалістів. Недоліком даного методу є швидкість перевірки повідомлень. Інформація в мережі Інтернет швидко розповсюджується, користувачі діляться нею, надсилаючи один одному повідомлення, тому доти, доки повідомлення буде перевірене спеціалістами, його можуть прочитати сотні людей. Тому переважним є другий спосіб, що полягає в автоматизації визначення дезінформації з використанням інтелектуальних методів обробки природної мови.

1.2 Визначення медичної дезінформації онлайн

Медична дезінформація – це неправдива інформація, що суперечить існуючому медичному розумінню, або необ'єктивна інформація, котра охоплює та популяризує лише частину фактів.

Медична дезінформація в соціальних мережах являє собою потенційну загрозу громадській охороні здоров'я, але масштаб цієї проблеми доки не є точним і конкретних механізмів для боротьби з нею не розроблено.

Неправдиві новини в медицині привертають увагу читачів, адже вони дають надію на подолання хвороб і обіцяють прості та доступні засоби лікування. Реальні ж новини у свою чергу не виправдовують сподівання людей і часом навіть руйнують надію тих, хто на неї сподівався. Тому фейкові новини стають привабливими і спонукають все більшу кількість людей вірити в них [5].

Ідентифікувати неправдиві новини можна, якщо людина має деякий базовий рівень знань з медицини. Однак, більшість людей не мають відношення до даної сфери і віддають перевагу тій інформації, що активно поширюється в медіа і зокрема звучить від відомих людей, навіть не дивлячись на те, що вона може бути неправдивою.

Перевірити, чи правдива медична інформація на тому чи іншому ресурсі, можна не лише звернувшись на пряму до спеціаліста в даній області. На сьогоднішній день є сайти, на котрих спеціалісти консультують та відповідають на запитання в режимі онлайн, є сайти перевірки фактів, а також медичні довідники, де також можна дізнатися інформацію.

На рис.1-3 представлені сайти на яких розміщуються достовірні статті на різну тематику, в тому числі й медичну. Для перевірки статей чи тверджень на даних сайтах на достовірність на кожному з них є рядок пошуку за ключовими словами.

Сайт Snopes.com містить розділ «Fact Checks» з твердженнями, достовірність яких перевірена експертами, у тому числі й на медичну тематику [6].

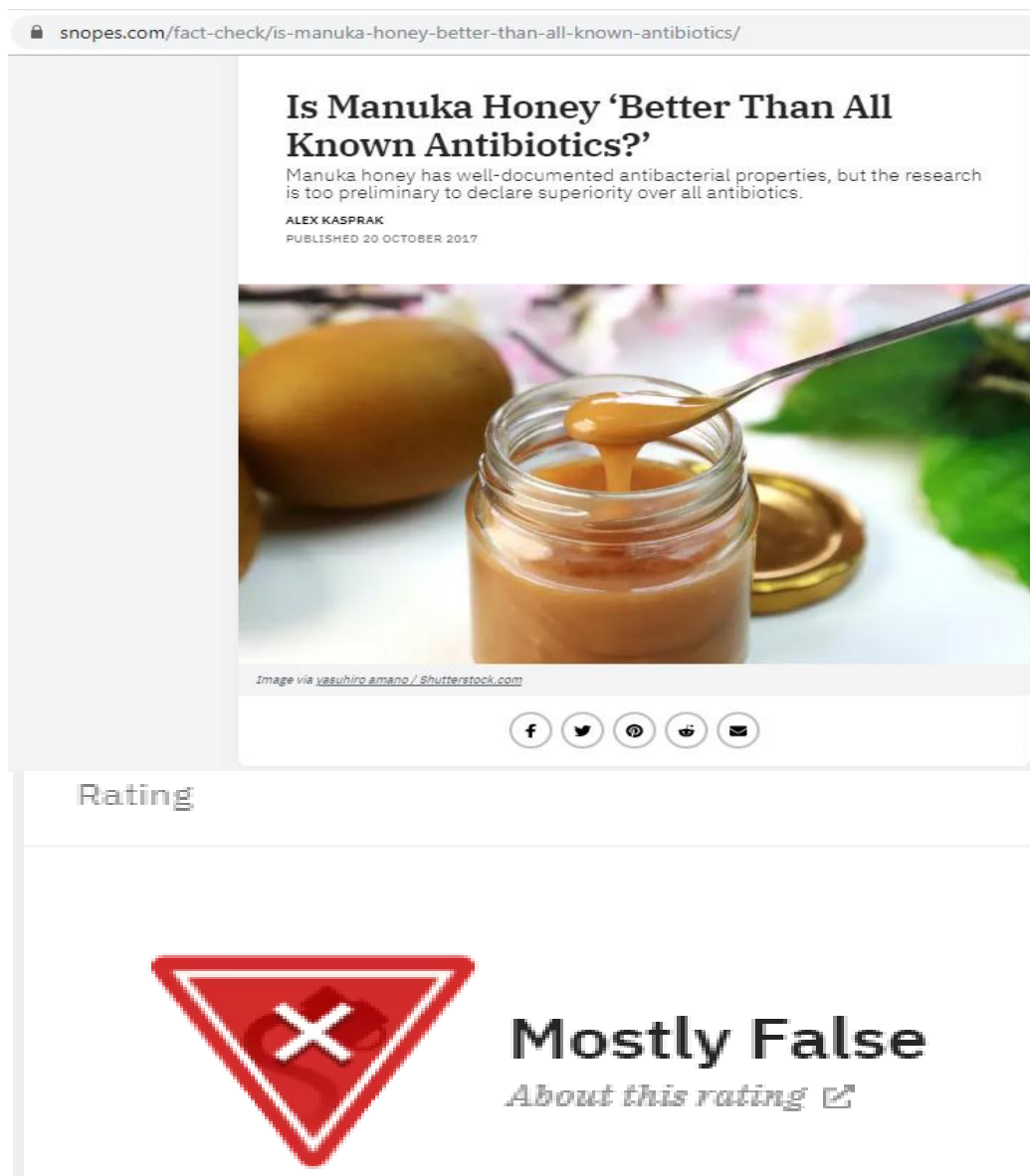


Рисунок 1.1 – Сайт з перевіреними статтями на достовірність

Платформа Metafact містить перевірені факти, у тому числі й на медичну тематику, на якій користувачі задають запитання, а профільні спеціалісти дають відповіді на них [7].

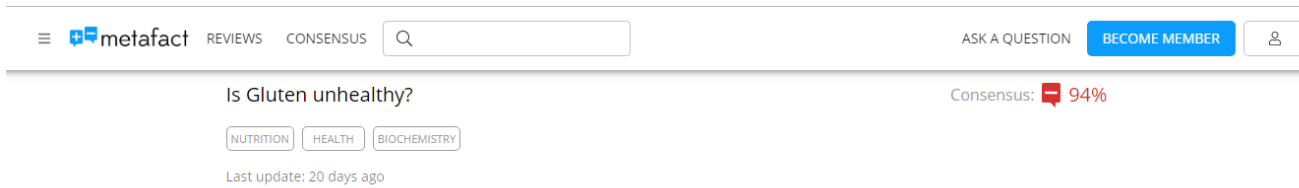


Рисунок 1.2 – Сайт з перевіреними статтями на достовірність

Сайт Health Feedback розміщує перевірені статті на медичну тематику, а також твердження [8].

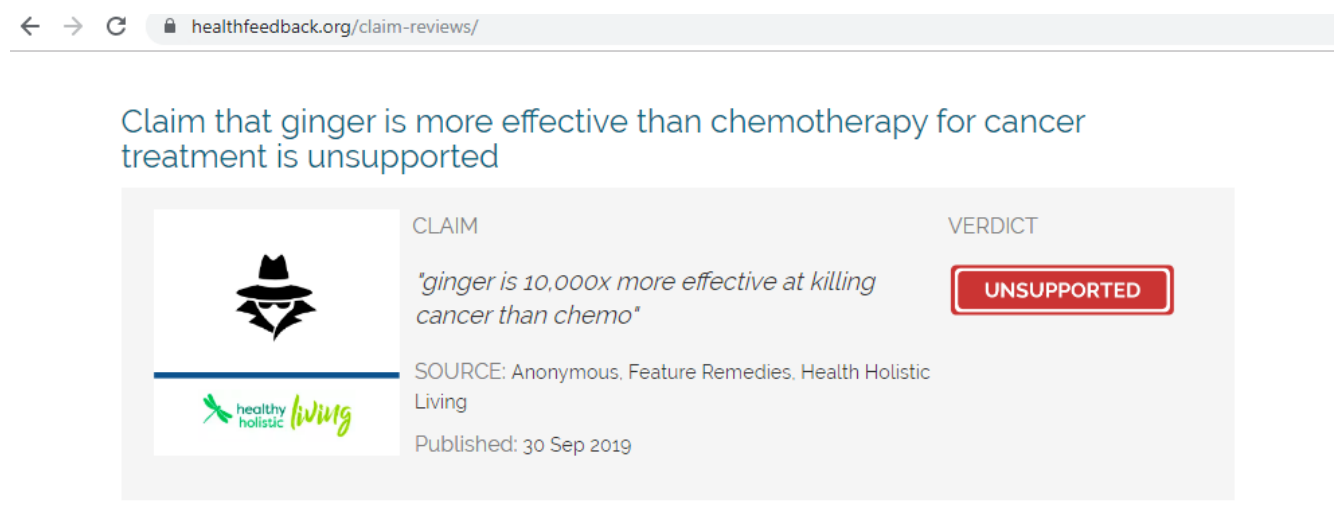


Рисунок 1.3 – Сайт з перевіреними статтями на достовірність

Недоліком є те, що всі дані необхідно перевіряти вручну, а це не завжди зручно зробити через брак часу чи відсутність доступу до мережі Інтернет.

1.3 Інтелектуальні методи обробки природної мови

Обробка природної мови (Natural Language Processing, NLP) – область, що знаходиться на перетині комп'ютерних наук, штучного інтелекту та лінгвістики. Мета полягає в обробці та «розумінні» природної мови для перекладу тексту і відповіді на питання [9].

З появою і бурхливим розвитком в мережі Інтернет текстової інформації, голосових інтерфейсів та чат-ботів значно прискорився розвиток NLP.

На даний час швидко зростає кількість корисних додатків, що розроблені з використанням методів обробки природної мови та вирішують задачі, такі як:

- пошук необхідної інформації;
- аналіз тональності тексту;
- показ онлайн реклами з урахуванням потреб та вподобань користувача;
- автоматичний переклад;
- аналіз задач маркетингу;
- розпізнавання мови (найчастіше використовується при розробці чат-ботів);
- голосові помічники.

Для того, щоб вирішити задачу усунення дезінформації в Інтернеті, необхідно використовувати механізм, котрий буде проводити аналіз цієї інформації та відфільтровувати її. Розробити його можна, використовуючи методи NLP.

Загалом виявлення дезінформації відбувається на різних рівнях, а саме – джерела, документа або твердження. Всі рівні базуються на різних методах виявлення дезінформації.

Методи, засновані на джерелах, як правило, прості і зосереджені на визначенні достовірності джерела інформації на основі попередньої поведінки представників джерела, наприклад, певного сайту новин, в мережі Інтернет. Визначення дезінформації на основі джерела є найпростішим, але найменш достовірним способом, тому що усі документи, розміщені від імені недостовірного джерела інформації, вважаються недостовірними.

Підходи, засновані на документах, перевіряють, чи містить документ дезінформацію. Інколи документ може містити як правдиву, так і неправдиву інформацію, тому його необхідно розбивати на більш дрібні частини і

використовувати методи, засновані на перевірці тверджень. Пошук дезінформації на рівні тверджень вимагає виконання декількох етапів, таких як витяг тверджень, поділ на речення і оцінка достовірності [10].

Більшість задач NLP починаються з попередньої обробки тексту, котра складається з декількох етапів:

- поділ на речення;
- лексемизація;
- нормалізація(лематизація);
- видалення стоп-слів.

Для вирішення задач, пов'язаних з обробкою природної мови, використовують ряд методів:

- наївний байєсівський класифікатор;
- метод опорних векторів;
- нейронні мережі;
- технологія Word2Vec;
- метод, заснований на правилах та словниках.

Часто в дослідженнях використовують методи на основі машинного навчання з учителем. Суть даних методів полягає в тому, що спочатку навчається машинний класифікатор на заздалегідь розмічених текстах, а потім отриману модель використовують при аналізі нових документів. Одним з методів навчання з учителем є наївний байєсівський класифікатор.

Наївний байєсівський класифікатор часто використовується в задачах класифікації текстів, при фільтрації спаму або визначення тональності документа. Перевагою наївного байєсівського класифікатора є невелика кількість даних для навчання, необхідних для оцінки параметрів, що враховуються при класифікації. Вхідними даними для реалізації наївного байєсівського класифікатора є навчальна вибірка, в якій визначено відповідність між текстовими документами та класами, до яких вони відносяться [11].

Метод опорних векторів (SVM) – це метод навчання з учителем, що використовується для бінарної класифікації. У даному методі кожний документ відображається у вигляді вектору в багатовимірному просторі. SVM на основі навчальних даних створює модель, яка визначає цільове значення тексту, та буде оптимальну розмежувальну гіперплощину. Метод опорних векторів досить стійкий до великих розмірностей даних, тому текстові дані підходять для використання класифікатора SVM [12].

Для визначення емоційного забарвлення тексту також можна використовувати нейронні мережі. Штучні нейронні мережі – це система з'єднаних і взаємодіючих між собою простих процесів, а саме – штучних нейронів. Алгоритм роботи таких процесів зазвичай простий, але якщо об'єднати їх у велику мережу з керованою взаємодією, можна виконувати досить складні завдання. Для вирішення задач NLP використовуються рекурентні та згорткові нейронні мережі [12].

В основі технології Word2Vec слова представляються у вигляді векторів заданої розмірності, розташовуючи схожі слова близько один до одного. Тобто, відстань між векторами слів, що позначають схожі речі, наприклад, "кіт" і "собака", буде значно менше, ніж між словами, значення яких мають мало спільного, наприклад, "кіт" і "літак". Дана особливість дозволяє більш гнучко представляти дані, які в подальшому можуть бути використані в навчанні нейронних мереж та різних класифікаторів. Для створення бази відповідностей "слово - вектор", алгоритм спочатку переглядає весь виданий йому текст, складаючи "словник", який в наступних ітераціях роботи алгоритму, буде використаний для визначення відповідних векторів [12].

Метод, заснований на правилах та словниках, виконує пошук емотивної лексики в тексті по заздалегідь складеними словниками та правилами із застосуванням лінгвістичного аналізу. Текст оцінюється за шкалою, що містить кількість негативної та позитивної лексики [12].

Альтернативою методам машинного навчання є використання підходу Elasticsearch. Elasticsearch широко використовується в якості пошукового і аналітичного апарату. Його можливості в якості API для аналізу текстової інформації не так відомі, але в нього переваги при вирішенні задач, пов'язаних з обробкою тексту. Він простий у використанні та швидкий. За допомогою даного підходу задача класифікації вирішується набагато швидше, ніж при використанні традиційного підходу. Але деякі задачі NLP потребують більш глибокого аналізу, і в цьому випадку Elasticsearch не має ідеальної архітектури та формату даних, і необхідно додатково використовувати плагіни, щоб мати доступ до тексту в цілому.

Описані вище методи можна також використовувати для вирішення задач, що стосуються фільтрації фейкових новин в Інтернеті, у тому числі й для виявлення медичної дезінформації, а саме для класифікації новин, тобто автоматично відрізнити фальшиві новини від справжніх; для припинення розповсюдження неправдивої інформації зрозумілим, швидким та дієвим способом.

2 ПОСТАНОВКА ЗАДАЧІ ТА МЕТОДИ ДОСЛІДЖЕННЯ

2.1 Мета та задачі

Метою даної роботи є розробка програмного додатка для визначення дезінформації на медичну тематику, що публікується на форумах. Тобто необхідно, використовуючи інтелектуальні методи аналізу природної мови, розробити додаток, котрий буде проводити аналіз текстової інформації, що зберігається в постах та коментарях на форумі, відфільтровувати дані, розділяючи їх на достовірні та неправдиві.

Для розробки даного додатка необхідно виконати ряд поставлених задач. Розроблена система призначена для модерації повідомлень на форумі, як постів так і коментарів. Тому однією із задач є організація обміну даними між форумом та програмним додатком, використовуючи API інтерфейс. Потрібно відслідковувати появу нових постів та коментарів та перевіряти їх відразу після публікації. Після отримання публікацій необхідно провести попередню обробку тексту, а саме видалити пробіли та цифри, тобто підготувати текст для подальшої роботи з ним.

Наступна задача, яку повинен виконати додаток після отримання публікацій та обробки тексту – виконати задачу класифікації, котра полягає у наступному.

Нехай ми маємо статтю a . Вона складається із двох основних компонентів: автор та контент. Автор P_a включає в себе набір атрибутів, таких як ім'я, вік і т.д. Зміст S_a складається з набору атрибутів статті і включає в себе адреса статті на форумі, заголовок, текст, зображення.

Визначимо взаємодію на форумі як набір кортежів $\mathcal{E} = \{e_{it}\}$ для представлення поширення публікацій в часі серед n користувачів $U = \{u_1, u_2, \dots, u_n\}$ і їх відповідних повідомлень (коментарів) $P = \{p_1, p_2, \dots, p_n\}$ в онлайн-спільноті щодо статті a . Кожна взаємодія $e_{it} = \{u_i, p_i, t\}$ означає, що користувач u_i коментує

статтю a , використовуючи ri за час t . Якщо $t = 0$, то значить стаття a ще не має коментарів і тоді $u_i \in \Pi$ її автором [13].

Отже, задача виявлення неправдивих новин на форумі полягає в тому, щоб визначити, чи містить публікація неправдиву інформацію, тобто $F: \mathcal{E} \rightarrow \{0,1\}$, а саме

$$F(a) = \begin{cases} 1, & \text{якщо новина фальшива} \\ 0, & \text{в іншому випадку} \end{cases} \quad (1)$$

де F - функція прогнозування.

Основними вимогами до додатку є те, що додаток повинен визначати дезінформацію в автоматизованому режимі, тобто без безпосередньої участі модератора. Додаток для пошуку дезінформації повинен бути реалізований англійською мовою, тому що більшість сайтів, що містять перевірену медичну інформацію, реалізовані англійською.

2.2 Огляд python-фреймворків для обробки природної мови

Python – високорівнева мова програмування загального призначення, орієнтована на підвищення продуктивності розробника і читання коду. Стандартна бібліотека включає великий обсяг корисних функцій. Python підтримує:

- структурне програмування;
- об'єктно-орієнтоване;
- функціональне;
- імперативне;
- аспектно-орієнтоване програмування.

Основні архітектурні риси мови Python – динамічна типізація, автоматичне керування пам'яттю, механізм обробки виключень, підтримка багатопоточних обчислень, високорівневі структури даних.

Для обробки природної мови в Python використовуються такі структури даних як словники, списки, кортежі та множини.

Python – мова програмування, що активно розвивається, нові версії з додаванням, зміною мовних властивостей виходять приблизно раз в два з половиною роки. Мова не піддавалася офіційній стандартизації, роль стандарту виконує CPython, що розробляється під контролем автора мови. На даний момент Python займає третє місце в рейтингу ТЮВЕ з показником 8,5%. На сьогоднішній день актуальною є версія Python 3.8.0 [14].

Python має ряд потужних бібліотек, котрі використовуються для вирішення різного виду задач і для обробки природної мови також. При розробці програмного додатка нам як раз і знадобляться бібліотеки Python для Natural Language Processing.

NLTK (Natural Language ToolKit) – пакет бібліотек і програм для символного і статистичної обробки природної мови. Підтримує роботу з великою кількістю мов. [15].

Переваги:

- найвідоміша і повна по функціоналу бібліотека для NLP;
- велика кількість сторонніх розширень;
- швидка токенізація речень;
- підтримується безліч мов.

Недоліки:

- повільна;
- складна в вивченні і використанні;
- працює з рядками;
- не використовує нейронні мережі;
- відсутні вбудовані вектори слів.

SpaCy – бібліотека, написана на мові Python, позиціонується як найшвидша NLP бібліотека. Має безліч можливостей, в тому числі, розбір залежностей на основі міток, розпізнавання іменованих сутностей, позначення частин мови, вектори розстановки слів [15].

Переваги:

- найшвидша бібліотека для NLP;
- проста у вивченні і використанні;
- працює з об'єктами, а не рядками;
- є вбудовані вектори слів;
- використовує нейронні мережі для тренування моделей.

Недоліки:

- менш гнучка в порівнянні з NLTK;
- токенизація пропозицій повільніше, ніж в NLTK;
- підтримує невелику кількість мов.

Scikit-learn – бібліотека, що надає реалізацію цілого ряду алгоритмів для навчання з учителем і навчання без учителя через інтерфейс для Python. Побудована поверх SciPy. Орієнтована в першу чергу на моделювання даних, має досить функцій, щоб використовуватися для NLP в зв'язці з іншими бібліотеками [15].

Переваги:

- велика кількість алгоритмів для побудови моделей;
- містить функції для роботи з Bag-of-Words моделлю;
- якісна документація.

Недоліки:

- попередня обробка змушує використовувати її в зв'язці з іншою бібліотекою (наприклад, NLTK);
- не використовує нейронні мережі для препроцесінга тексту.

Gensim – Python бібліотека для тематичного моделювання документів і обробки природної мови. У gensim реалізовані популярні NLP алгоритми, наприклад, word2vec. Більшість реалізацій можуть використовувати кілька ядер [15].

Переваги:

- працює з великими наборами даних;
- підтримує глибоке навчання;
- word2vec, tf-idf vectorization, document2vec.

Недоліки:

- працює з моделями без вчителя;
- не містить достатнього функціоналу, необхідного для NLP, що змушує використовувати її разом з іншими бібліотеками.

Кожна з бібліотек має свої переваги та недоліки, і яку саме обрати залежить вже від поставленої задачі, котру необхідно вирішити.

Також для вирішення задач пов'язаних з NLP, Python має фреймверки машинного навчання такі як PyTorch, Keras та TensorFlow.

PyTorch - це середовище машинного навчання на мові Python з відкритим вихідним кодом, що забезпечує тензорні обчислення з GPU-прискоренням. Використовується для вирішення різних задач: комп'ютерне бачення, обробка природної мови [16].

Переваги:

- має велику кількість модульних елементів, які легко комбінувати.
- легко писати власні типи шарів і працювати на GPU.
- має широкий вибір попередньо навчених моделей

Недоліки:

- тренувальний код доведеться писати самостійно.
- Недостатня кількість документації.

Keras - відкрите нейромережеве середовище глибокого навчання, написане на Python. Фреймворк містить численні реалізації широко вживаних блоків нейронних мереж, таких як шари, цільові та передавальні функції, оптимізатори, та безліч інструментів для спрощення роботи із зображеннями та текстом [16].

Переваги:

- зручний у використанні;
- легкий в засвоєнні;
- вбудований в TensorFlow.
- якісна документація.

Недоліки:

- не підходить для великих проектів.

TensorFlow - це комплексна платформа для машинного навчання з відкритим вихідним кодом. TensorFlow помітно спрощує вбудовування в додатки елементів, що самонавчаються, і функцій штучного інтелекту, призначених для розпізнавання мови, організації комп'ютерного зору або обробки природної мови [16].

Переваги:

- відмінний фреймворк для створення нейронних мереж;
- бере на себе оптимізацію ресурсів для обчислень;

Недоліки:

- складний у використанні і освоєнні;
- недружній;
- необхідно постійно контролювати використовувану відеопам'ять;
- має свої стандарти;
- недостатня кількість документації.

Для розробки було обрано такі бібліотеки як pandas, numpy, та keras. Адже вони найбільш підходять для задач, котрі необхідно вирішити.

2.3 Засоби реалізації програмного продукту

Розробка програмного додатку відбувалася за допомогою середовища Anaconda.

Anaconda - дистрибутив мов програмування Python і R, що включає набір популярних вільних бібліотек, об'єднаних проблематиками науки про дані і машинного навчання. Anaconda містить єдиний узгоджений комплект найбільш використовуваних тематичних модулів (таких як NumPy, SciPy, Astropy і інших) з вирішенням залежностей і конфліктів, які неминучі при одиночній установці даних модулів. Станом на 2019 рік містить більше 1,5 тис. модулів.

Основна особливість дистрибутива - оригінальний менеджер Anaconda Navigator, що дозволяє відмовитися від стандартних менеджерів пакетів (таких, як рір для Python). Дистрибутив потрібно завантажити один раз, і вся подальша конфігурація, в тому числі установка додаткових модулів, може проводитися в офлайн. Крім того, забезпечується можливість ведення декількох ізольованих середовищ [17].

Всі необхідні бібліотеки та фреймверки наявні в Anaconda Navigator, їх лише необхідно підключити.

Однією із задач при розробці програмного продукту є витяг публікацій із форуму для подальшої роботи з ними. Так я форум було розроблено за допомогою системи Wordpress, тому і витягуються публікації з використанням WP REST API. Даний інтерфейс повноцінно доданий до ядра Wordpress в версії 4.7. Він дає можливість користувачам віддалено взаємодіяти із сайтом: відправляти запити і отримувати відповіді. Відповіді надходять у JSON форматі. Так, наприклад, можна отримувати пости, рубрики, коментарі з сайту. А також крім цього такі ж дані можна створювати, змінювати та видаляти з сайту. Відбувається все це через URL запити так звані «маршрути» [18].

2.3.1 Формування набору даних для класифікації тверджень

Вхідний набір даних , що використовується для навчання мережі було взято із платформи Monant. Дана платформа - це універсальна та розширювана платформа, розроблена для моніторингу сайтів та витягу з них публікацій, у тому числі на медичну тематику [19].

Отримані дані з платформи записуються до json файлу, а потім конвертуються у файл формату csv для подальшої роботи з ними рис.2.1-2.2.

Імпортуємо модуль Requests, бібліотеки роботи з json, csv та Pandas

Задаємо адресу доступу до сторінки авторизації на платформі

```
url = 'https://api.monant.fiit.stuba.sk/auth'
```

Задаємо параметри для пошукового рядку

```
payload = "{\n    \"username\": \"parfenenko\",\n    \"password\": \"33ef38b5-181f-4b3f-ba00-e2479b8bb0db\"\n}"
```

вказуємо в заголовку, що будемо працювати з json

```
headers = {\n    'Content-Type': 'application/json'}
```

Відправляємо запит зі встановленими параметрами на сервер, якщо дані вказані вірно, то відбувається авторизація

```
generatedtoken = requests.request('POST', url, headers = headers, data = payload)
```

перетворюємо результат запиту у json

```
a=generatedtoken.json()
```

Після авторизації звертаємося на сервер також через API для того, щоб одержати твердження на тематику.

```
url2 = ' https://api.monant.fiit.stuba.sk/v1/claims?'
```

Формуємо пошуковий рядок

```
category=medical&search=cancer&rating=false&page=1'
```

Вказуємо у заголовку вид аутентифікації

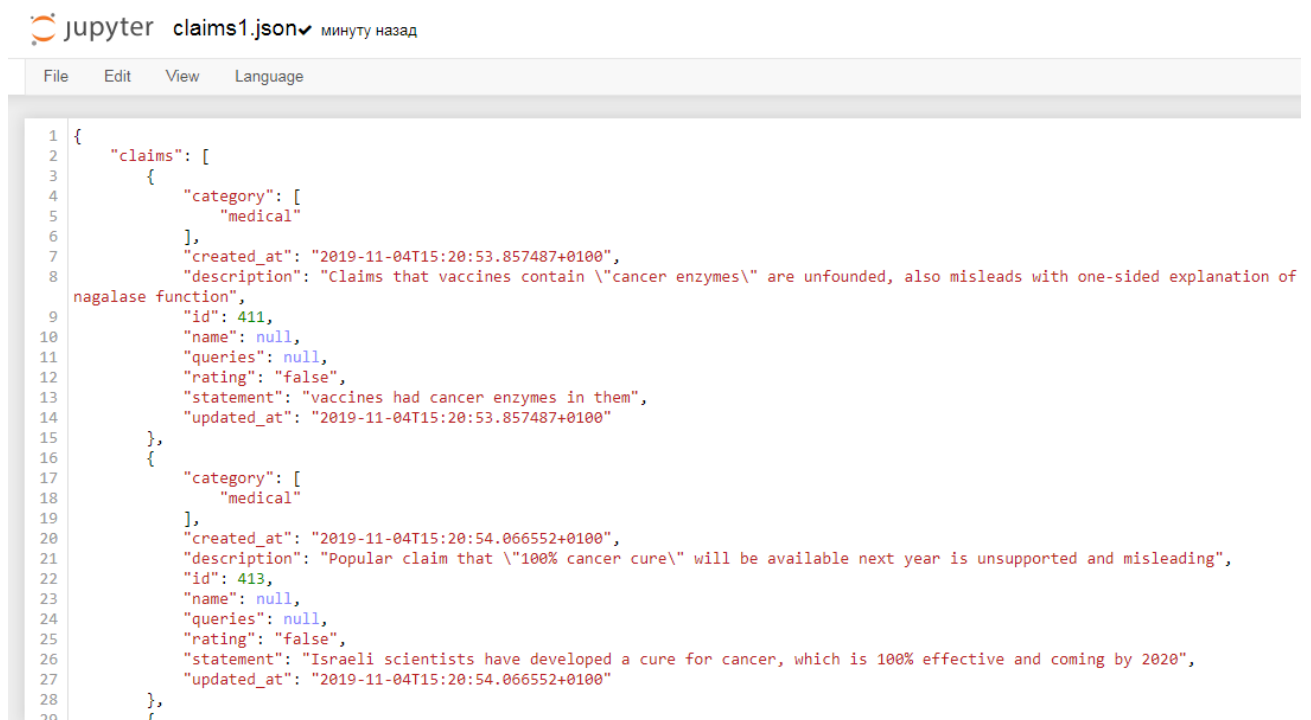
```
headers = {
    'Authorization': 'JWT ' + a['access_token'],
    'Content-Type': 'application/json'}
```

Виконуємо запит на отримання даних

```
response = requests.request('GET', url2, headers = headers, data = payload,
allow_redirects=False)
```

Створюємо файл json та записуємо в нього результат запиту

```
todos = json.loads(response.text)
with open('claims1.json', 'a', encoding='utf-8') as f:
    json.dump(todos, f, ensure_ascii=False, indent=4)
```



```
jupyter claims1.json ✓ минуту назад
File Edit View Language
1 {
2   "claims": [
3     {
4       "category": [
5         "medical"
6       ],
7       "created_at": "2019-11-04T15:20:53.857487+0100",
8       "description": "Claims that vaccines contain \"cancer enzymes\" are unfounded, also misleads with one-sided explanation of
nagalase function",
9       "id": 411,
10      "name": null,
11      "queries": null,
12      "rating": "false",
13      "statement": "vaccines had cancer enzymes in them",
14      "updated_at": "2019-11-04T15:20:53.857487+0100"
15    },
16    {
17      "category": [
18        "medical"
19      ],
20      "created_at": "2019-11-04T15:20:54.066552+0100",
21      "description": "Popular claim that \"100% cancer cure\" will be available next year is unsupported and misleading",
22      "id": 413,
23      "name": null,
24      "queries": null,
25      "rating": "false",
26      "statement": "Israeli scientists have developed a cure for cancer, which is 100% effective and coming by 2020",
27      "updated_at": "2019-11-04T15:20:54.066552+0100"
28    },
29  ]
30 }
```

Рисунок 2.1 – Файл з твердженнями з платформи Monant

Як бачимо твердження включають в себе різні поля, такі як:

Category – категорія до якої відноситься твердження. Created_at – дата створення твердження. Description – опис твердження. Id – ідентифікатор джерела

статті. Name – назва. Queries – запити. Rating – рейтинг твердження: true – правдиве, false хибне. Statement – саме твердження. Update_at – дата оновлення.

Для формування вхідного набору даних використовуються поля такі як Statement та Rating.

Витяг даних було здійснено на такі медичні теми як: анемія, діабет, вакцини, рак та грип, а потім їх було поєднано в єдиному файлі.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	rating,"statement"															
2	false,"Automobile components emit dangerous levels of cancer-causing benzene fumes."															
3	false,"Hiding finely-ground glass in food is an effective method for killing a person."															
4	false,"Cancer is caused by a deficiency of vitamin "B17," a condition that can be remedied with nutritional supplements."															
5	false,"College student who restricts himself to only one foodstuff contracts scurvy."															
6	false,"Ordinary use of canola oil is dangerous to consumers."															
7	true,"Some brands of sidewalk chalk have contained high levels of lead."															
8	false,"vaccines had cancer enzymes in them"															
9	false,"side effects of the [HPV] vaccine outweigh the positives"															
10	false,"Each of these ingredients [thimerosal, ovalbumin, formaldehyde, sodium deoxycholate] is a known neurotoxin"															
11	false,"vaccines cause autism"															
12	false,"With repeated use of vaccines, cell-mediated immunity weakens"															
13	false,"If a child has a mitochondrial disorder [...] she is more at risk of having a vaccine pushing her over the edge and causing a bad reaction"															
14	false,"Measles Vaccines Kill More People than Measles, CDC Data Proves"															
15	false,"Toxic vaccine ingredients in the MMR vaccine cause encephalopathy"															
16	false,"Formerly healthy people are suffering serious harm or even dying from taking the flu shot"															
17	true,"HPV vaccine may lead to elimination of cervical cancer"															
18	true,"HPV vaccine safe"															
19	true,"Flu Vaccine effective"															
20	true,"A virus deadly to dogs has been spreading in the canine population of the U.S."															
21	false,"Ingesting ""a tablespoon"" of fruit syrup every 15 minutes for an hour can help someone dealing with vomiting or diarrhea."															
22	false,"Artificially sweetened foods and drinks are less healthy, than their sugary counterparts."															
23	false,"Tur as a result, India, whose food contains a large amount of the spice, has the lowest rates of dementia worldwide."															
24	false,"Diets tailored to one's specific blood type are capable of reducing myriad ailments, improving digestion, enabling weight loss, and providing increased energy."															

Рисунок 2.2 – Вхідний набір даних для класифікації тверджень

У результаті одержуємо розмічений набір даних, який може бути використаний для бінарної класифікації тверджень.

2.4 Модель нейронної мережі для класифікації тверджень

Визначення медичної дезінформації буде відбуватися з використанням нейронної мережі, а саме згорткової нейронної мережі (convolutional neural network, CNN). CNN – це клас глибоких штучних нейронних мереж прямого поширення. Широкого застосування згорткової нейронної мережі набули при

класифікації зображень [20]. На даний час CNN застосовуються для задач NLP і показують високу ефективність. Тому що цей тип може використовуватися для роботи з текстом, для його класифікації також.

Архітектура згорткової нейронної мережі для класифікації тексту зображена на рис.2.3.

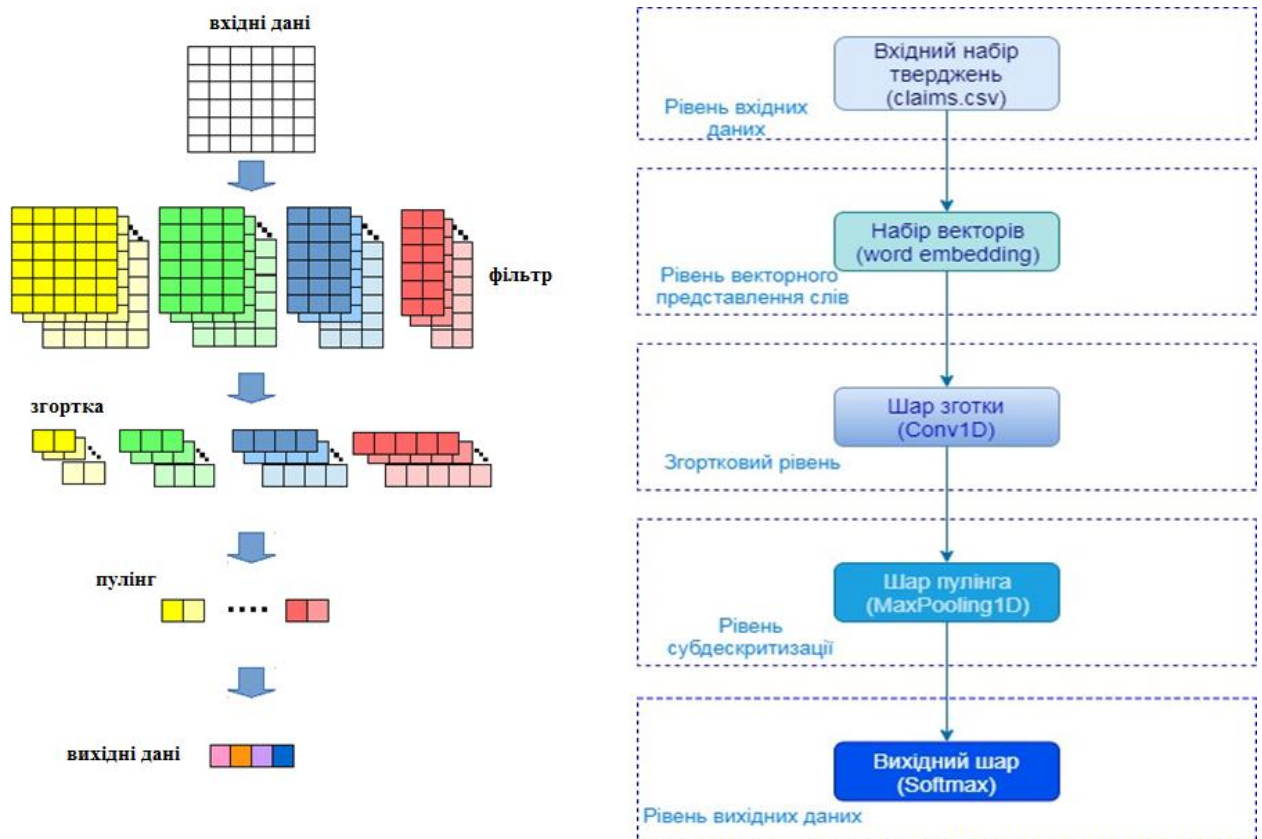


Рисунок 2.3 – Архітектура згорткової нейронної мережі

Нейронна мережа створювалася мовою Python з використанням бібліотеки Keras. Для створення нейронної мережі CNN у Keras потрібно викликати функцію з такими параметрами.

На вхід поступає вхідний набір тверджень, отриманий з платформи Monant. Набір тверджень спочатку розбивається на навчальну та тестову вибірки. Далі твердження проходять попередню обробку: видалення зайвих символів, пробілів та розбиваються на слова.

Так як нейронна мережа працює з числовими даними, текст представляється у вигляді числових векторів, за допомогою моделі word2vec. Сформовані числові

вектори подаються на одновимірний згортковий шар Conv1D, де відбувається процес згортки. Результат згортки поступає на рівень субдескрипції MaxPooling1D, де формується підвибірка. Після цього сформована вибірка подається на повністю пов'язану нейронну структуру, яка приймає рішення про класифікацію виходячи з ваг, присвоєних кожній ознаці в тексті.

Перш ніж модель буде готова до навчання, нам потрібно вказати ще кілька параметрів: функція втрат (loss function), оптимізатор (optimizer), метрики(metrics). Вони додаються на етапі compile моделі.

Навчання моделі відбувається за допомогою методу model.fit, котрий тренує модель на тренувальних даних.

В процесі навчання моделі відображаються метрики втрати (loss) і точності (accuracy).

Наша модель досягає на тренувальних даних точності дорівнює приблизно 0.92 (92%) рис.2.4.

```

Epoch 7/20
109/109 [=====] - 0s 3ms/step - loss: 0.0625 - acc: 0.9908 - val_loss: 1.1211 - val_acc: 0.7500
Epoch 8/20
109/109 [=====] - 0s 3ms/step - loss: 0.0380 - acc: 1.0000 - val_loss: 1.0569 - val_acc: 0.7500
Epoch 9/20
109/109 [=====] - 0s 3ms/step - loss: 0.0384 - acc: 1.0000 - val_loss: 1.0000 - val_acc: 0.7500
Epoch 10/20
109/109 [=====] - 0s 3ms/step - loss: 0.0305 - acc: 1.0000 - val_loss: 0.9563 - val_acc: 0.7500
Epoch 11/20
109/109 [=====] - 0s 4ms/step - loss: 0.0300 - acc: 1.0000 - val_loss: 0.9307 - val_acc: 0.7500
Epoch 12/20
109/109 [=====] - 0s 4ms/step - loss: 0.0281 - acc: 1.0000 - val_loss: 0.9123 - val_acc: 0.8333
Epoch 13/20
109/109 [=====] - 0s 4ms/step - loss: 0.0304 - acc: 1.0000 - val_loss: 0.9021 - val_acc: 0.8333
Epoch 14/20
109/109 [=====] - 0s 4ms/step - loss: 0.0239 - acc: 1.0000 - val_loss: 0.8925 - val_acc: 0.8333
Epoch 15/20
109/109 [=====] - 0s 3ms/step - loss: 0.0194 - acc: 1.0000 - val_loss: 0.8859 - val_acc: 0.8333
Epoch 16/20
109/109 [=====] - 0s 4ms/step - loss: 0.0211 - acc: 1.0000 - val_loss: 0.8793 - val_acc: 0.8333
Epoch 17/20
109/109 [=====] - 0s 5ms/step - loss: 0.0184 - acc: 1.0000 - val_loss: 0.8727 - val_acc: 0.8333
Epoch 18/20
109/109 [=====] - 0s 3ms/step - loss: 0.0200 - acc: 1.0000 - val_loss: 0.8661 - val_acc: 0.9167
Epoch 19/20
109/109 [=====] - 0s 3ms/step - loss: 0.0203 - acc: 1.0000 - val_loss: 0.8592 - val_acc: 0.9167
Epoch 20/20
109/109 [=====] - 0s 3ms/step - loss: 0.0183 - acc: 1.0000 - val_loss: 0.8528 - val_acc: 0.9167

```

Рисунок 2.4 – Навчання моделі

Далі, було порівняно, яку точність модель показує на тестовому датасеті рис.2.5.

```
In [31]: test_loss, test_acc = model.evaluate(X_val, y_val)
print( "Test accuracy:", test_acc)

12/12 [=====] - 0s 3ms/step
Test accuracy: 0.9166666865348816
```

Рисунок 2.5 – Точність навчання та тестовому наборі даних

Для того щоб отримати якомога більшу точність було проведено декілька експериментів, під час яких відбувалася зміна параметрів нейронної мережі (табл.2.1).

Таблиця 2.1 – Підбір параметрів нейронної мережі.

№	К-ть навчальних зразків (batch_size)	К-ть епох (num_epochs)	L2_регулізація (l2_lambda)	К-ть фільтрів (filters)	Розмір підвибірки (pool_size)	Розмір ядра (kernel_size)	К-ть моделей (ens_models)	Точність (accuracy)
1	64	3	0.001	100	2	3	3	0.69
2	32	20	0.0001	32	2	3	3	0.75
3	64	10	0.0001	32	2	3	3	0.80
4	128	30	0.0001	32	2	3	3	0.83
5	128	20	0.0001	32	2	3	2	0.916

Вхідний набір даних необхідно розділити на декілька навчальних зразків. Параметр «кількість навчальних зразків» - це загальна кількість тренувальних об'єктів, представлених в одному пакеті. Параметр «к-ть епох» - це кількість ітерацій навчального алгоритму по всій навчальній множині. L2_регулізація знижує складність моделі, зберігши кількість її параметрів. Параметр «к-ть фільтрів» відповідає за кількість ядер в згорткових шарах. В шарі субдискретизації параметр «pool_size» задає розмір підвибірки. За розмір ядра в згорткових шарах відповідає параметр «розмір ядра». Параметр «к-ть моделей» вказує число моделей, котрі будуть навчені за допомогою вхідної вибірки.

Підібравши найбільш оптимальні параметри, було проведено навчання мережі точність якого складає 0.916.

Таким чином, було створено нейронну мережу, яка буде використовуватися для бінарної класифікації тверджень при визначення медичної дезінформації.

3 МОДЕЛЮВАННЯ ПРОГРАМНОГО ДОДАТКУ ВИЗНАЧЕННЯ МЕДИЧНОЇ ДЕЗІНФОРМАЦІЇ В ОНЛАЙН-СПІЛЬНОТАХ

3.1 Структурно-функціональне моделювання процесу визначення медичної дезінформації в онлайн-спільнотах

Методології та стандарти дозволяють описати і наочно уявити як саме працює той чи інший продукт та з яких процесів складається його робота. Для повного розуміння роботи програмного продукту зручно використовувати структурно-функціональне моделювання в нотації IDEF0 та IDEF3.

За допомогою графічної мови IDEF0 розроблюваний програмний додаток постає у вигляді набору взаємопов'язаних функціональних блоків, що значно полегшує розуміння всіх процесів, котрі він виконує. За допомогою IDEF3 описується логіка виконання дій. IDEF3 може використовуватися спільно з методологією IDEF0 і самостійно: будь-який функціональний блок IDEF0 може бути представлений у вигляді послідовності процесів або операцій засобами IDEF3. IDEF0 описує процеси роботи продукту, а IDEF3 описує як ці процеси виконуються [21].

Визначення медичної дезінформації відбувається на форумі, запит на перевірку публікації поступає на вхід автоматично з появою нової публікації. Обмежують даний процес документація по роботі з API, правила попередньої обробки тексту, нейромережеві методи обробки природної мови та вимоги до точності визначення дезінформації. Відбувається визначення медичної дезінформації за участі апаратного забезпечення, API для роботи з Wordpress та Monant, модулю витягу даних та модулю визначення дезінформації, а також методу word2vec. На виході ми отримуємо результат перевірки публікацій, а саме – текст публікації в кабінеті адміністратора форуму, в якій відмічено кольором твердження, які були визначені програмним додатком як неправдиві.

На рис.3.1 зображена контекстна діаграма визначення медичної дезінформації в онлайн-спільнотах.

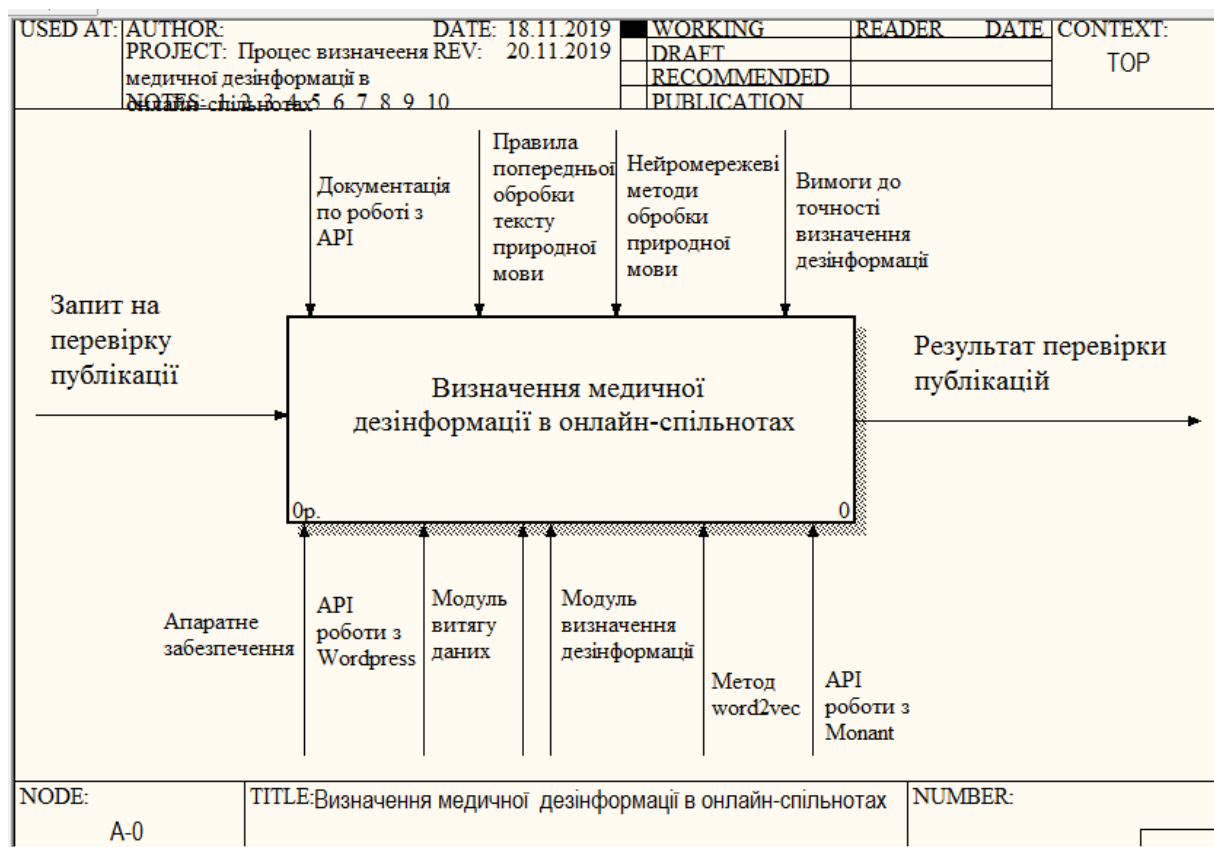


Рисунок 3.1 - Контекстна діаграма у нотації IDEF0

Проведено декомпозицію контекстної діаграми на функціональні блоки «Отримання публікації», «Обробка тексту», «Класифікація тверджень», «Візуалізація результату».

Декомпозиція 1-го рівня моделювання роботи програмного додатку визначення медичної дезінформації представлена на рис.3.2.

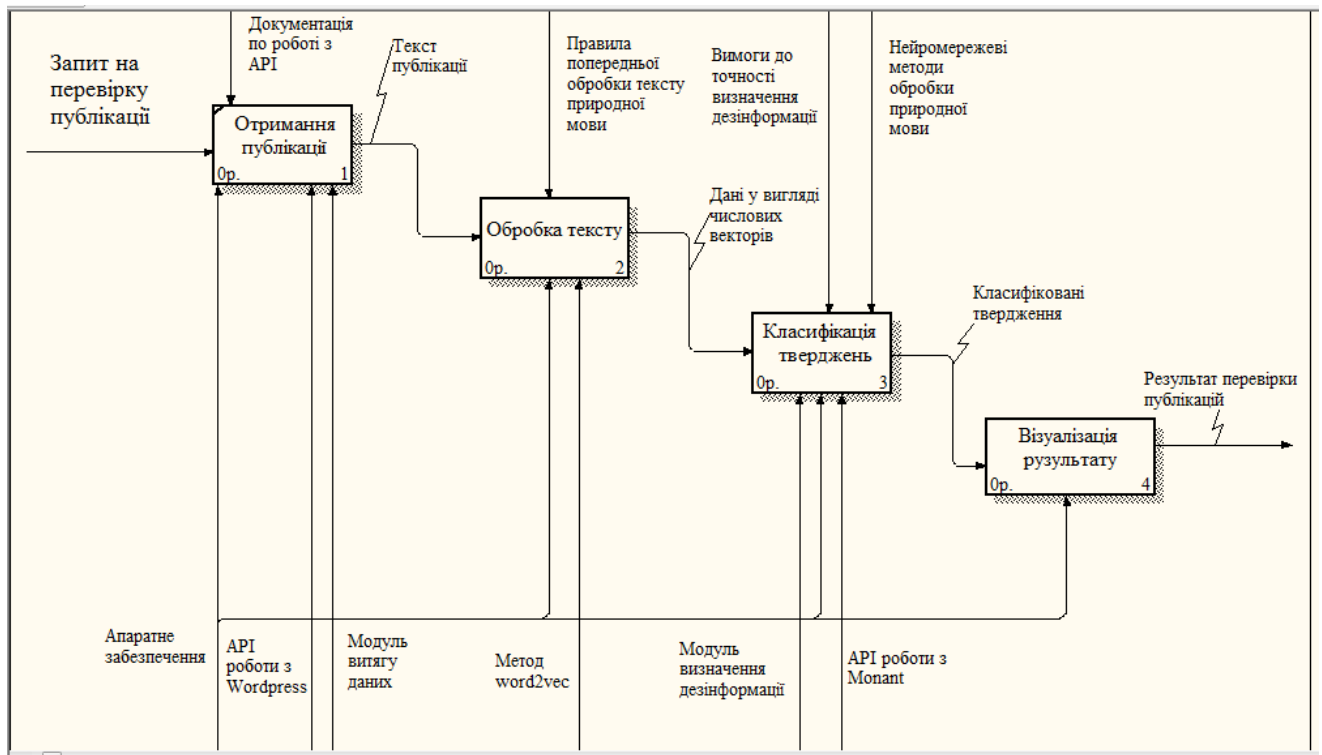


Рисунок 3.1 – Діаграма декомпозиції 1-го рівня у нотації IDEF0

На першому процесі необхідно за допомогою API інтерфейсу WordPress отримати публікацію, що знаходиться на форумі. Публікація складається із статті та коментарів до неї. На наступному етапі потрібно провести обробку тексту, а саме провести очистку тексту від тегів, розбити його на масив тверджень, провести токенизацію та перетворити текст на числові вектори для подальшої роботи.

Декомпозиція 2 –го рівня процесу «Обробка тексту» представлена у вигляді IDEF3 діаграми на рис.3.3.

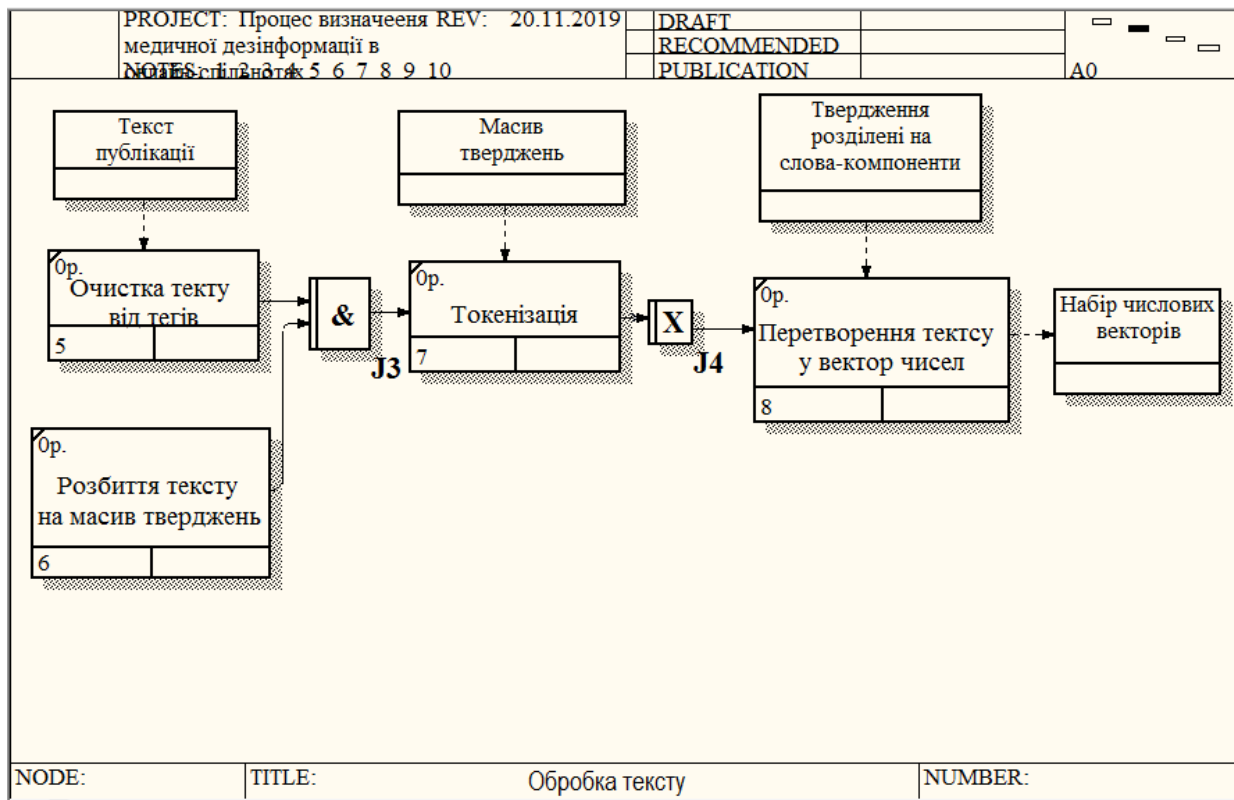


Рисунок 3.3 - Декомпозиція 2 –го рівня етапу обробки тексту

Виконання третього процесу полягає у класифікації тверджень за допомогою нейронної мережі. На даному етапі відбувається витяг тренувальних даних за допомогою API-платформи Monant. Підготовка тренувального набору даних для нейронної мережі складається з тих самих кроків, що і підготовка публікацій. Далі відбувається побудова та навчання нейронної мережі з використанням тренувального набору даних. Наступний крок полягає в оцінці точності моделі, після чого відбувається використання моделі та отримання класифікованих тверджень.

Декомпозиція 2 –го рівня процесу «Класифікація тверджень» представлена у вигляді IDEF3 діаграми на рис.3.4.

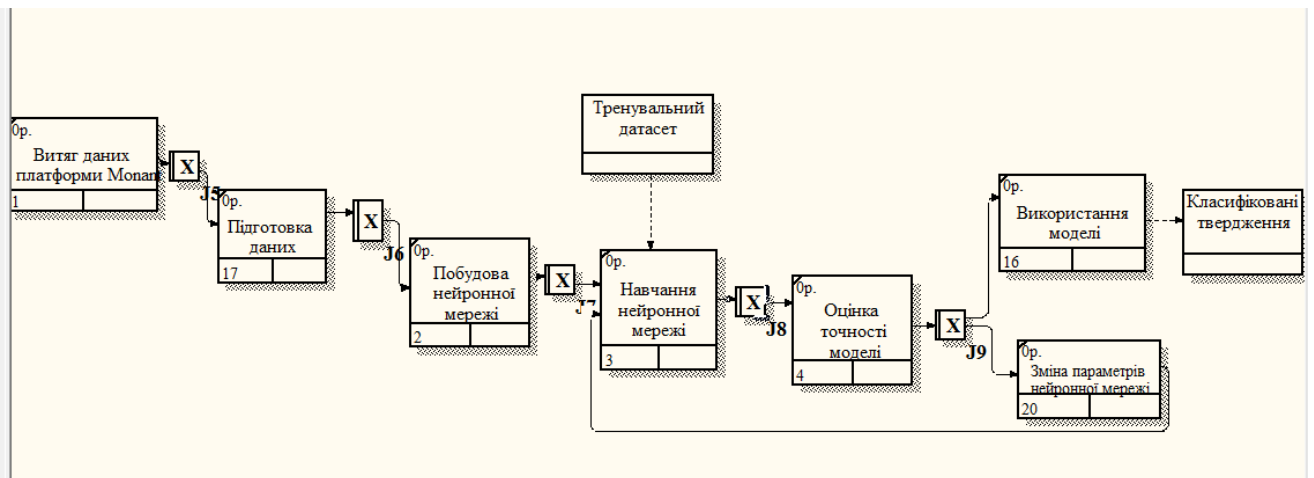


Рисунок 3.4 - Декомпозиція 2 –го рівня етапу класифікації тверджень

Етап візуалізації результатів – це розробка кабінету адміністратора на форумі. В даному кабінеті буде відображатися перевірена публікація і речення, що містять дезінформацію будуть відмічені певним кольором, щоб наглядно показати, де саме знаходиться неправдива інформація.

3.2 Побудова діаграми варіантів використання

Для опису взаємодії між об'єктами та програмним продуктом, було розроблено діаграму варіантів використання.

Діаграма прецедентів (діаграма варіантів використання) — діаграма, на якій зображено відношення між акторами та прецедентами в системі.

Діаграма прецедентів є графом, що складається з множини акторів, прецедентів (варіантів використання) обмежених границею системи (прямокутник), асоціацій між акторами та прецедентами, відношень серед прецедентів, та відношень узагальнення між акторами. Діаграми прецедентів відображають елементи моделі варіантів використання [22]

Діаграма варіантів використання програмного додатку визначення медичної дезінформації зображена на рис.3.5.

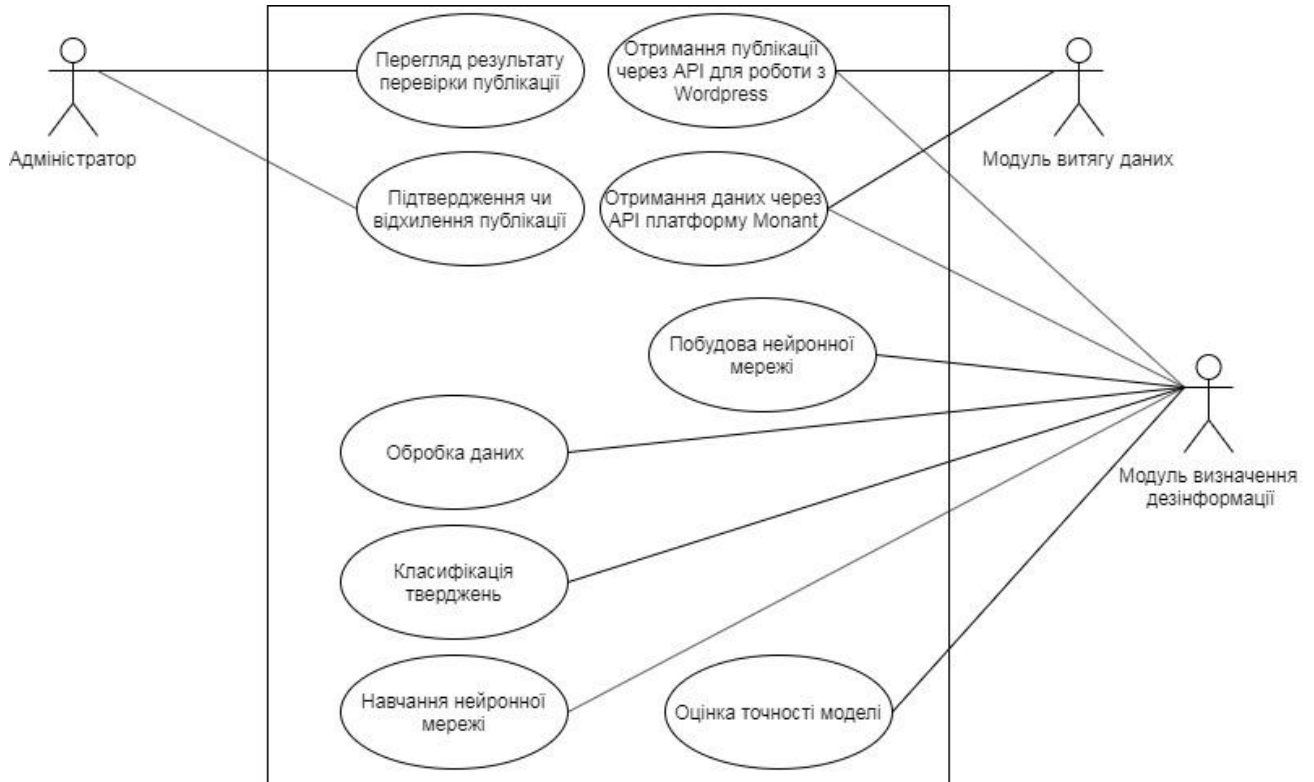


Рисунок 3.5 – Діаграма варіантів використання

На діаграмі варіантів використання визначено три актори: адміністратор–адміністратор форуму, котрий переглядає результат перевірки публікації та підтверджує чи відхилює публікацію; модуль витягу даних за допомогою якого відбувається отримання даних з форуму та платформи Monant; модуль визначення дезінформації ,що відповідає за обробку даних, створення та навчання мережі, класифікацію тверджень та оцінку точності .

3.3 Побудова діаграми діяльності

Діаграма діяльності – це діаграма, яка показує, як потік управління переходить від однієї діяльності до іншої, при цьому увага фіксується на результаті діяльності. Результат може привести до зміни стану системи або повернення деякого значення. Графічно діаграма діяльності представляється у формі графа діяльності, вершинами якого є стан дії або стан діяльності, а дугами - переходи від одного стану дії / діяльності до іншого. Кожна діаграма діяльності повинна мати єдиний початковий і єдиний кінцевий стан [23].

В якості графічного представлення для виділення основних функцій додатку було використано діаграму варіантів використання (use case). Діаграма варіантів використання дає нам представлення, що повинен робити додаток, а діаграма послідовності показує, як саме він виконує певні дії [24].

На рис.3.6 представлено діаграму діяльності для модулю визначення медичної дезінформації, де показано весь шлях твердження від публікації на форумі до перевіреного твердження.



Рисунок 3.6 – Діаграма діяльності модулю визначення медичної дезінформації

Результатом діяльності модулю визначення медичної дезінформації є класифіковані твердження.

3.4 Логічне проектування моделі бази даних

Логічне проектування — це створення схеми бази даних на основі уже існуючої моделі даних [25]. У даній роботі база даних створюється автоматично при встановленні CMS Wordpress. В ній зберігаються дані про користувачів форуму, публікації, коментарі, категорії. У таблиці 3.1 наведено список основних таблиць БД.

Таблиця 3.1 – Таблиці БД

Назва таблиці	Опис
Wp_users	Дані про користувачів форуму. Їх логін, пароль, статус.
Wp_posts	Дані про публікації. Вказується їх автор, дата створення, заголовок, контент, статус публікації.
Wp_comments	Дані про коментарі. Їх автор, контент, id публікації.
Wp_terms	Категорії публікацій.

На рис.3.7 приведена логічна модель бази даних

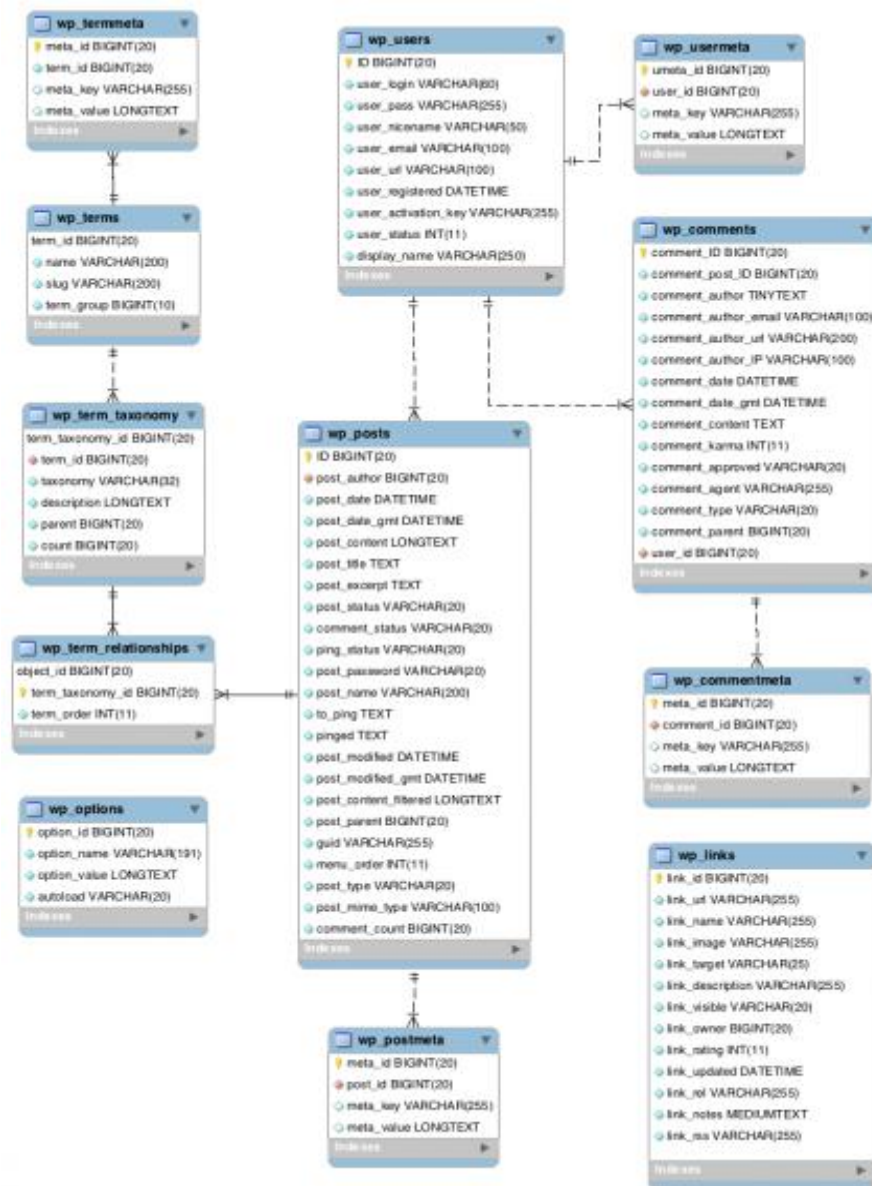


Рисунок 3.7 - Схема бази даних

Програмний додаток працює з таблицею `wp_posts`, де зберігаються публікації. З цієї таблиці додаток бере публікації, де в полі `post_status` вказано статус «pending». Результат перевірки також заноситься до даної таблиці в поле `post_check`. Якщо перевіряються коментарі на наявність дезінформації, то додаток працює з таблицею `wp_comments`.

4 РЕАЛІЗАЦІЯ ПРОГРАМНОГО ДОДАТКУ ВИЗНАЧЕННЯ МЕДИЧНОЇ ДЕЗІНФОРМАЦІЇ

4.1 Архітектура роботи програмного додатку

Архітектура - це основна організація системи, що втілюється в її компонентах, їх відносинах один з одним та з навколишнім середовищем, а також принципах, що визначають дизайн та розвиток системи [26].

Архітектура програмного додатку представлена на рис.4.1 у вигляді діаграми компонентів.

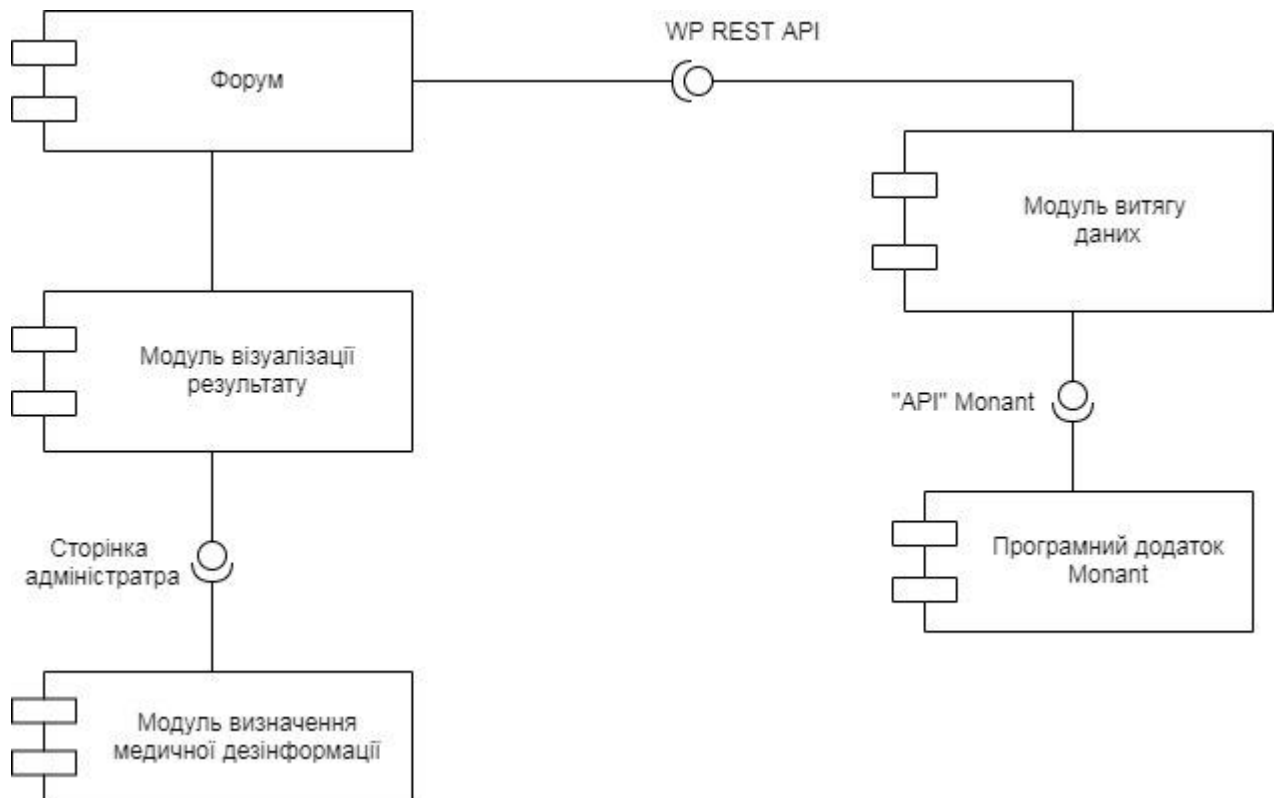


Рисунок 4.1 – Архітектура програмного додатку визначення медичної дезінформації

Діаграма компонентів - діаграма, на якій відображаються компоненти, залежності та зв'язки між ними. Вона відображає залежності між компонентами програмного забезпечення, включаючи компоненти вихідних кодів, бінарні

компоненти, та компоненти, що можуть виконуватись. Модуль програмного забезпечення може бути представлено як компоненту [27].

Для тестування додатку було створено форум на WordPress, з котрим програмний додаток безпосередньо взаємодіє і підтримує його адміністрування.

Публікації витягуються з форуму, за допомогою модулю витягу даних та WP REST API. Також даний модуль витягує твердження для класифікації з платформи Monant, використовуючи API інтерфейс. Після отримання всіх даних вони передаються до модулю визначення медичної дезінформації, де на основі навченої нейронної мережі відбувається класифікація. Після цього результат класифікації відображається у модулі візуалізації результатів, а саме в кабінеті адміністратора форуму.

4.2 Результати роботи

Щоб перевірити роботу розробленого додатку, виконати його тестування було розроблено форум. Форум був розроблений з використанням CMS Wordpress рис.4.2. В якості оформлення підключено тему, що найбільш відповідає тематиці форуму, додано віджети, меню та наповнення головної сторінки.



Рисунок 4.2 – Головна сторінка форуму

На рис.4.3 зображено сторінку форуму, де розміщуються публікації, зокрема й на медичну тематику.

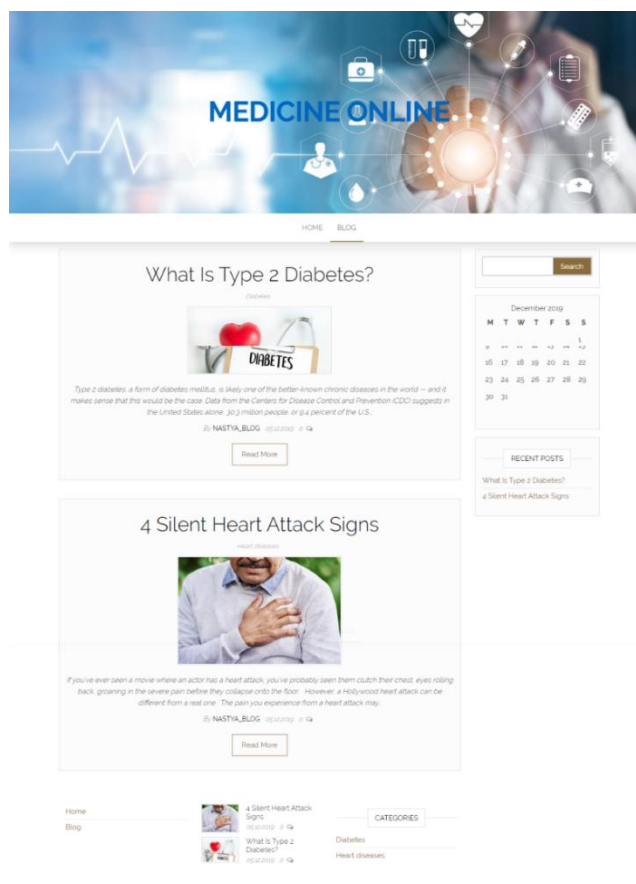


Рисунок 4.3 – Сторінка публікацій форуму

Розмістити публікацію можна, використовуючи спеціальну форму для додавання нових публікацій, зображену на рис.4.4



The image shows a forum page with a header banner for 'MEDICINE ONLINE' featuring medical icons and a doctor's hand holding a glowing orb. Below the banner are navigation links for 'HOME' and 'BLOG'. The main content area is titled 'Add post' and contains a form with the following fields:

- Your Name:** Peter
- Post Title:** Cancer Enzymes in Vaccines
- Post Category:** Cancer, Vaccine
- Text:** American doctors have discovered the enzyme nivalase, which is considered a cancerous enzyme in vaccines. Nagalase suppresses the immune system and therefore will be bad if found in vaccines. There have been controversial studies that claim to show that the GcMAF protein, which aims to lower the nagalase
- Submit Post** button

The sidebar on the right includes a search bar, a calendar for December 2019, and a 'RECENT POSTS' section listing 'What Is Type 2 Diabetes?' and '4 Silent Heart Attack Signs'.

Рисунок 4.4 – Додавання публікації на форум

Перед розміщенням публікації на форумі вона повинна пройти перевірку на наявність у ній медичної дезінформації. На сторінці адміністратора відображається нова публікація, яку необхідно перевірити. Натиснувши кнопку «Text check», запускається модуль визначення дезінформації рис.4.5.



HOME BLOG

Protected: Admin Page

New posts(1)

Cancer Enzymes in Vaccines

SHOCK! SHOCK!

American doctors have discovered the enzyme nivalase, which is considered a cancerous enzyme in vaccines. Nagalase suppresses the immune system and therefore will be bad if found in vaccines. There have been controversial studies that claim to show that the GcMAF protein, which aims to lower the nagalase concentration in the body, can treat diseases like cancer, autism, or HIV. These are all diseases in which the level of nagalase should be high.

[TEXT CHECK](#)

HOME BLOG

December 2019						
M	T	W	T	F	S	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

RECENT POSTS

What Is Type 2 Diabetes?

4 Silent Heart Attack Signs

Рисунок 4.5 – Сторінка адміністратора з неперевіреними публікаціями

Публікація має два статуси опублікована(publish), тобто вже перевірена додатком і статус очікує перевірки(pending).

Спочатку неперевірений пост витягується з бази даних, за допомогою WP REST API та записується до csv файлу рис.4.6.

```

jupyter posts.csv  минуто назад
File Edit View Language current mode
2 Cancer Enzymes in Vaccines,"<p>SHOCK! SHOCK!<br />
3 American doctors have discovered the enzyme nivalase, which is considered a cancerous enzyme in<br />
4 vaccines.<br />
5 Nagalase suppresses the immune system and therefore will be bad if found in vaccines. There have been controversial studies that claim to
show that the GcMAF protein, which aims to lower the nagalase concentration in the body, can treat diseases like cancer, autism, or HIV.
These are all diseases in which the level of nagalase should be high.</p>
6
7
8
9
10
11
12
13
14

```

Рисунок 4.6 – Файл csv з неперевіреною публікацією

Далі проводиться класифікація тверджень за допомогою модулю визначення медичної дезінформації. Текст постів розбивається на речення та дані речення класифікуються за допомогою навченої моделі.

Результат класифікації зберігається до бази даних. В таблиці з публікаціями в полі `post_check` вказується статус перевіреного твердження: правдиве/неправдиве рис.4.7.

<input type="checkbox"/>	8	post_status	varchar(20)	utf8mb4_unicode_ci	Нет	publish
<input type="checkbox"/>	9	post_check	tinyint(1)		Нет	Нет
<input type="checkbox"/>	10	comment_status	varchar(20)	utf8mb4_unicode_ci	Нет	open
<input type="checkbox"/>	11	ping_status	varchar(20)	utf8mb4_unicode_ci	Нет	open
<input type="checkbox"/>	12	post_password	varchar(255)	utf8mb4_unicode_ci	Нет	
<input type="checkbox"/>	13	post_name	varchar(200)	utf8mb4_unicode_ci	Нет	
<input type="checkbox"/>	14	to_ping	text	utf8mb4_unicode_ci	Нет	Нет

Рисунок 4.7 – Поля, де відображається статус поста.

Нейронна мережа перевіряє кожне твердження, та визначає % точності класифікації. В базу даних заноситься точність класифікації кожного твердження і якщо воно неправдиве, його колір змінюється в залежності від % точності. Відображається результат на сторінці адміністратора рис. 4.8.

HOME BLOG

Protected: Admin page

New posts(1)

Cancer Enzymes in Vaccines

SHOCK! SHOCK!

American doctors have discovered the enzyme nagalase, which is considered a cancerous enzyme in vaccines.

Nagalase suppresses the immune system and therefore will be bad if found in vaccines. There have been controversial studies that claim to show that the GcMAF protein, which aims to lower the nagalase concentration in the body, can treat diseases like cancer, autism, or HIV. These are all diseases in which the level of nagalase should be high.

DELETE POST

December 2019

M	T	W	T	F	S	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

Рисунок. 4.8 – Відображення результату перевірки публікації в кабінеті адміністратора

Якщо в тексті публікації виявлено дезінформацію на медичну тематику, вона виділяється кольором з загального тексту поста, і його адміністратор має змогу видалити.

ВИСНОВКИ

Через стрімкий розвиток мережі Інтернет та великий потік інформації дезінформація в мережі Інтернет розповсюджується з великою швидкістю. Вона має негативний вплив на окремих користувачів та суспільство в цілому. Виявлення неправдивих новин, у даний час, є актуальною проблемою, яка на даний момент не має конкретного алгоритму розв'язання.

Даний програмний додаток був розроблений у рамках кваліфікаційної роботи. На початку роботи було визначено мету та сформовано основні задачі, котрі необхідно вирішити. Було проведено аналіз предметної області та огляд основних методів, що використовуються обробки природної мови.

У якості методу реалізації задачі було обрано метод нейромережевої класифікації. Сформовано набір даних для класифікації.

Побудовано згорткову нейронну мережу, проведено обчислювальні експерименти з вибору її параметрів.

Під час виконання роботи було проведено моделювання роботи програмного додатку, за допомогою діаграм у нотації IDEF та UML діаграм.

Результатом роботи є програмний додаток визначення медичної дезінформації в онлайн-спільнотах. Даний додаток протестовано на прикладі розробленого форуму. Він допомагає адміністратору знайти в публікації неправдиву інформацію, та недопустити розміщення цієї інформації на форумі.

СПИСОК ЛІТЕРАТУРИ

1. Антисоциальное поведение : веб=сайт. URL: <https://psyera.ru/antisocialnoe-povedenie-612.htm> – (дата звернення:18.10.2019).
2. Srijan Kumar, Justin Cheng, Jure Laskovec. Antisocial Behavior on the Web: Characterization and Detection. Proceedings of the 26th International Conference on World Wide Web Companion ,2017. P. 947-950.
3. Фальшивые новости: URL: https://ru.wikipedia.org/wiki/Фальшивые_новости – (дата звернення:18.10.2019).
4. Shivam B. Parikh, Pradeep K. Atrey. Media-Rich Fake News Detection: A Survey. Albany Lab for Privacy and Security, College of Engineering and Applied Sciences University at Albany, State University of New York, Albany, NY, USA. 2017. P. 33-38.
5. Правда про фейкові новини в медицині URL: <http://ingeniusua.com/uk/node/2881> – (дата звернення:18.10.2019).
6. Snopes.com: веб=сайт. URL: <https://www.snopes.com/> – (дата звернення:24.10.2019).
7. Metafact: веб=сайт. URL: <https://metafact.io/> – (доступ з екрану – 24.10.2019).
8. Health Feedback: веб=сайт. URL: <https://healthfeedback.org/> – (дата звернення:24.10.2019).
9. 5 методов обработки естественного языка, которые стремительно меняют мир вокруг нас : веб=сайт. URL: <https://neurohive.io/ru/osnovy-data-science/5-metodov-v-nlp-kotorye-izmenjat-obshhenie-v-budushhem/> – (дата звернення:25.10.2019).
- 10.Co-Inform: Co-Creating Misinformation-Resilient Societies / T. Farrell, M. Mensio, Gr. Burel, L. Piccolo, H. Alani. 2018. P. 58

- 11.Б. О. Близнюк, Л. В. Васильева, И. Д. Стрельников, Д. С. Ткачук. Современные методы обработки естественного языка. Вісник Харківського національного університету імені В. Н. Каразіна. 2017. Вип. 36. С. 14-26.
- 12.Аналіз тональності тексту: веб-сайт. URL: https://uk.wikipedia.org/wiki/Аналіз_тональності_тексту – (дата звернення:18.10.2019).
- 13.Kai Shu , Amy Sliva , Suhang Wang , Jiliang Tang , Huan Liu. Fake News Detection on Social Media: A Data Mining Perspective. *Computer Science & Engineering, Arizona State University, Tempe, AZ, USA*. 2017. P. 15.
- 14.Python:веб-сайт. URL: <https://ru.wikipedia.org/wiki/Python>
<https://ru.wikipedia.org/wiki/Python> – (дата звернення:27.10.2019).
- 15.Обработка естественного языка. URL: http://neerc.ifmo.ru/wiki/index.php?title=Обработка_естественного_языка – (дата звернення:27.10.2019).
- 16.Топ-10 фреймворков для искусственного интеллекта: часть первая: URL: <https://vc.ru/ml/80391-top-10-freymvorkov-dlya-iskusstvennogo-intellekta-chast-pervaya> – (дата звернення:20.11.2019).
- 17.Anaconda: URL: <https://ru.wikipedia.org/wiki/Anaconda> – (дата звернення:20.11.2019).
- 18.WP REST API: URL: <https://wp-kama.ru/handbook/rest> - (дата звернення:18.11.2019).
- 19.Ivan Srba, Robert Moro, Jakub Simko, Jakub Sevcech, Daniela Chuda, Pavol Navrat, Maria Bielikova. Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. Slovak University of Technology Bratislava, Slovakia.P.1-7.
- 20.Згорткова нейронна мережа: URL: https://uk.wikipedia.org/wiki/Згорткова_нейронна_мережа - (дата звернення:25.11.2019)

21. Stud.com.ua: веб=сайт. URL: https://stud.com.ua/97325/informatika/metodologiyi_standarti_funktsionalnogo_modelyuvannya_predmetnoyi_oblasti – (дата звернення: 18.11.2019).
22. Діаграми Прецедентів (Use Case UML Diagram) URL: <http://lvivqaclub.blogspot.com/2008/10/use-case-uml-diagram.html> - (дата звернення: 18.11.2019).
23. Діаграма діяльності: URL: <http://studepedia.org/index.php?vol=1&post=2126> - (дата звернення: 25.11.2019)
24. Теория и практика UML. Диаграмма деятельности: URL: http://it-gost.ru/articles/view_articles/96 - (дата звернення: 25.11.2019).
25. Проектування бази даних: URL: https://uk.wikipedia.org/wiki/Проектування_бази_даних - (дата звернення: 25.11.2019).
26. Створення програмного продукту: URL: <http://lib.mdpu.org.ua/e-book/vstup/L6.htm> - (дата звернення: 28.11.2019).
27. Діаграма компонентів: URL: https://uk.wikipedia.org/wiki/Діаграма_компонентів - (дата звернення: 05.12.2019).
28. Цели по SMART: подробный обзор: веб=сайт. URL: <http://powerbranding.ru/marketing-strategy/smart-celi/> – (дата звернення: 11.11.2019).
29. Управление проектами. Ру: веб=сайт. URL: <https://upravlenie-proektami.ru/chto-takoe-wbs-proekta-i-zachem-ona-nuzhna> – (дата звернення: 11.11.2019).
30. Finswin: веб=сайт. URL: <https://finswin.com/projects/instrumenty/kalendarnyj-plan-proekta.html> – (дата звернення: 11.11.2019).

ДОДАТОК А

Планування робіт з розроблення програмного додатку

4.3 Ідентифікація мети методом SMART

Для конкретизації мети магістерської роботи використаємо метод SMART. Метод SMART - сучасний підхід до постановки працюючих цілей. Він дозволяє на етапі визначення мети провести узагальнення всієї наявної інформації, встановити надати всім учасникам процесу ясні, точні, конкретні завдання.

SMART визначає критерії ефективності поставлених цілей. Розглянемо кожен критерій SMART-мети більш докладно [1]:

S – Specific (конкретна) – зазначається, яких результатів необхідно досягти.

M – Measurable (вимірювана) – вказується, в чому буде вимірюватися результат. Показник може бути кількісний або якісний.

A – Achievable (досяжна) – пояснюється, за рахунок чого планується досягти мети.

R – Relevant (значущість) – узгодженість з іншими завданнями. Визначається на скільки завдання релевантне, адекватне поточним зовнішнім і внутрішнім умовам.

T – Time-framed (обмежена у часі) – обмежена в часі мета означає наявність чітких термінів її досягнення.

У таблиці 2.2 представлено деталізацію мети кваліфікаційної роботи, що полягає у розробленні програмного додатку визначення медичної дезінформації в онлайн-спільнотах, з використанням методу SMART.

Таблиця А.1 - Деталізація мети методом SMART

S	Розробка програмного додатку виявлення медичної дезінформації в публікаціях на форумі
M	Результат буде вимірюватися % точно класифікованих тверджень у публікаціях
A	Для визначення медичної дезінформації буде використано нейромережеві методи обробки природної мови. Навчання нейронної мережі буде здійснюватися на наборі даних, які зібрані на платформі Monant. Для отримання даних з платформи Monant та публікацій з форуму система використовуватиме наявні API інтерфейси.
R	У наявності є всі необхідні технічні та програмні засоби. Знання розробників та консультантів дозволяють досягти поставленої мети.
T	Мета має часове обмеження. Робота має бути завершена до 10.12.2019 року, що є крайнім терміном здачі кваліфікаційної роботи магістра

Після проведення аналізу методом SMART можна визначити кінцеву мету.

Отже, мета проекту – розробка програмного додатку виявлення медичної дезінформації в онлайн-спільнотах, а саме додатку для адміністратора форуму на WordPress, за допомогою якого можна буде перевіряти публікації та коментарі до них на наявність медичної дезінформації без залучення особи, що є фахівцем з медицини. Це дозволить попередити швидке розповсюдження неправдивої медичної інформації та уникнути можливих наслідків самолікування серед відвідувачів форуму.

4.4 Планування змісту структури робіт ІТ-проекту

Виробнича структура (WBS - work breakdown structure) – це розбиття проекту на конкретні результати, які повинні бути досягнуті для досягнення цілей проекту. Вона є багаторівневою, кожний з рівнів формується в результаті розподілу роботи попереднього рівня на її складові. Як правило, на верхньому рівні вказується сам проект, під ним (на першому рівні) - основні роботи, кожен з яких, в свою чергу, деталізується, тобто наступний рівень завжди менше попереднього за обсягом робіт і, як правило, включає 2 і більше пакетів робіт. При цьому в різних гілках WBS може бути різна кількість рівнів в залежності від потрібного ступеня деталізації [25]. WBS діаграма декомпозиції робіт з розроблення програмного додатку визначення медичної дезінформації в онлайн-спільнотах представлена на рис. 1.2.

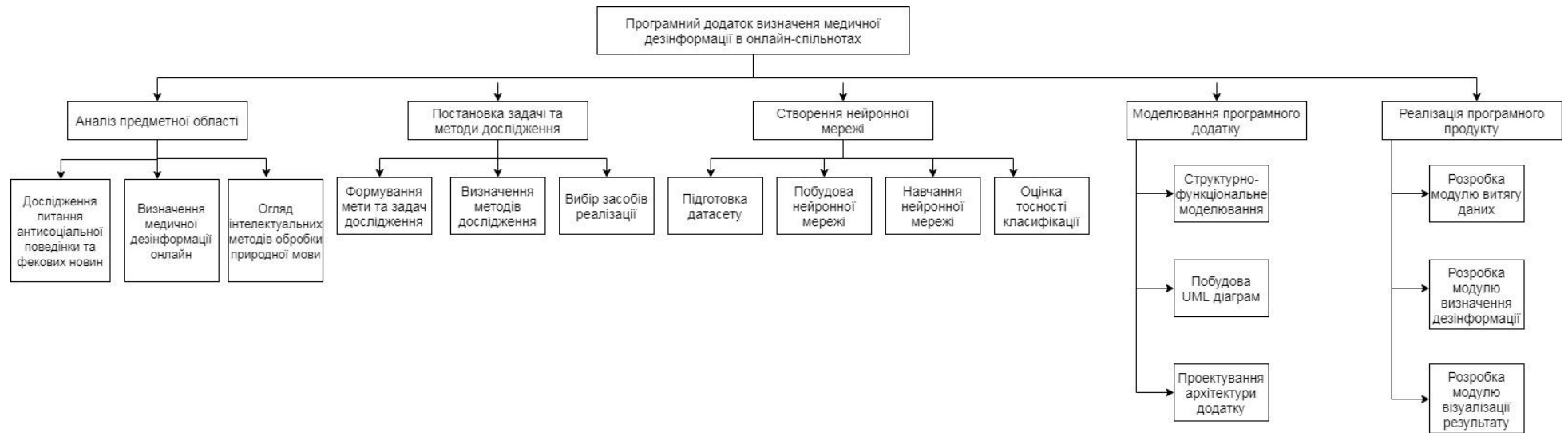


Рисунок А.1 – WBS діаграма декомпозиції робіт з розроблення програмного додатку визначення медичної дезінформації в онлайн-спільнотах

4.5 Побудова календарного графіку виконання ІТ-проекту

Для визначення часу, потрібного на виконання робіт з розроблення програмного додатку визначення медичної днзінформації в онлайн-спільнотах і для встановлення взаємозв'язків між ними з урахуванням найбільш ризикових подій, необхідно скласти календарний план проекту. Календарне планування полягає в створенні і подальшому уточненні розкладу, який враховує склад робіт, терміни їх виконання, ризики, обмеження [26].

Календарний графік робіт можна подати, використовуючи діаграму Ганта.

Діаграма Ганта представляє собою горизонтальну лінійну діаграму, на якій задачі проекту представляються часовими відрізками, що характеризуються датами початку та закінчення, а також затримками [27].

Кожен відрізок відповідає окремій задачі або підзадачі. Задачі і підзадачі розміщують по вертикалі. Початок, кінець і довжина відрізка на шкалі часу відповідають початку, кінцю і тривалості завдання.

На рисунку А.2 представлено діаграму Ганта розроблюваного проекту.

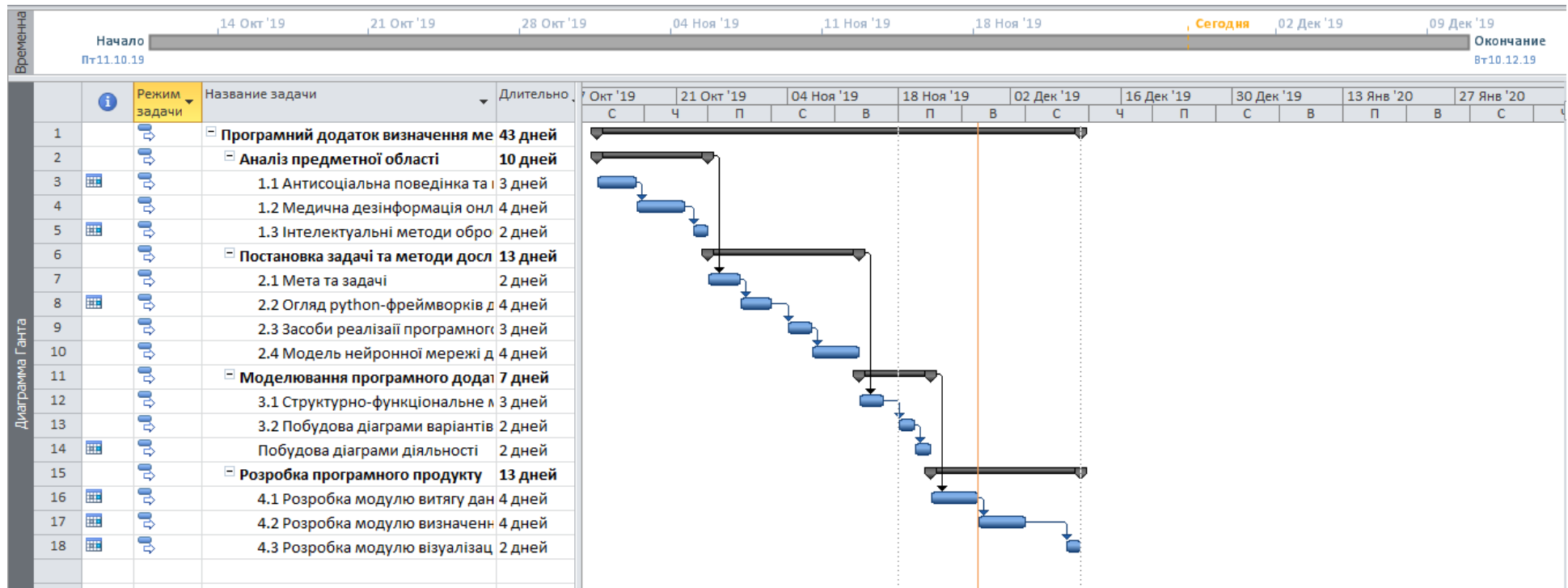


Рисунок А.2 – Діаграма Ганта виконання робіт з розроблення програмного продукту визначення медичної дезінформації в онлайн-спільнотах

4.6 Планування ризиків ІТ-проекту

Планування та реалізація проектів відбувається в умовах невизначеності, що породжується зміною внутрішнього та зовнішнього середовищ. Під невизначеністю розуміють відсутність повної та достовірної інформації про умови реалізації проекту.

Ризик - невизначена подія або безліч подій, які в разі реалізації вплинуть на досягнення цілей. Ризикове подія - це завжди «можливість» і «загроза».

Управління ризиком – це процес реагування на події та зміни ризиків в процесі виконання проекту.

При розробці будь-якого проекту необхідно передбачити всі можливі ризики, які можуть трапитись та провести необхідні заходи для їх уникнення чи зменшенню збитків.

Виділимо основні ризики під час розробки програмного продукту:

1. Поява альтернативного продукту.
2. Недотримання строків виконання проекту.
3. Вибір непідходящої технології.
4. Помилки в документації.
5. Помилки проектування.
6. Недостатність кваліфікації для реалізації поставлених задач.
7. Збій апаратного і програмного забезпечення.
8. Відсутність резервних копій даних.
9. Виникнення незапланованих робіт.

Для класифікації ризиків використаємо шкалу ймовірності виникнення та величину втрат (таблиця А.4.1).

Таблиця А.4.1 – Шкала оцінювання ймовірності виникнення та величини витрат

Оцінка	Ймовірність виникнення	Величина втрат
1	Слабоймовірна	Мінімальна
2	Малоймовірна	Низька
3	Ймовірна	Середня
4	Досить ймовірна	Висока
5	Майже можлива	Максимальна

Класифікуємо ризики згідно таблиці 2.1 та розрахуємо індекс ризику за формулою А.1

$$R = P_q * I_q \quad (\text{A.1})$$

де R – індекс ризику (бали);

P_q – ймовірність виникнення ризиків відповідно до класифікації (бали);

I_q – величина втрат відповідно до класифікації ризику (бали).

Таблиця А.4.2 – Оцінка ймовірності виникнення, величини витрат та індексу ризику

№	Ризики	P_q	I_q	R
1	Поява альтернативного продукту	2	3	6
2	Недотримання строків проекту	1	3	3
3	Вибір оптимальної технології	3	4	12
4	Помилки в документації	3	4	12
5	Помилки проектування	3	5	15
6	Недостатність кваліфікації для реалізації поставлених задач	4	4	16
7	Збій апаратного і програмного забезпечення	2	4	8
8	Відсутність резервних копій даних	1	4	4
9	Виникнення незапланованих робіт	3	3	9

Ймовірність	5	*	*	*	*	*
	4	*	*	*	9/10	*
	3	*	*	13	4/7	8
	2	*	5	1	11	3
	1	*	*	2	12	6
			1	2	3	4

Збитки

Позначення ризиків у матриці	
	Прийнятні
	Виправдані
	Недопустимі
	* Пусте значення

Рисунок А.4 – Матриця ризиків

На підставі отриманого значення індексу ризику класифікуємо: за ступенем впливу(таблиця А.4.3) та рівнем ризику(таблиця А.4.4).

Таблиця А.4.3 – Шкала оцінювання за ступенем впливу

№	Назва	Межі	Ризики, які входять
1	Ігноруючі	$1 \leq R \leq 4$	2. Недотримання строків проекту. 8. Відсутність резервних копій даних.
2	Незначні	$5 \leq R \leq 8$	1. Поява альтернативного продукту. 7. Збій апаратного і програмного забезпечення.
3	Помірні	$9 \leq R \leq 10$	9. Виникнення незапланованих робіт
4	Істотні	$12 \leq R \leq 16$	3. Вибір оптимальної технології. 4. Помилки в документації. 5. Помилки проектування. 6. Недостатність кваліфікації для реалізації поставлених задач
5	Критичні	$20 \leq R \leq 25$	Відсутні

Таблиця А.4.4 – Шкала оцінювання за рівнем ризику

№	Назва	Межі	Ризики, які входять(номера)
1	Прийнятні	$1 \leq R \leq 4$	2, 8
2	Виправдані	$5 \leq R \leq 10$	1, 7, 9
3	Неприпустимі	$12 \leq R \leq 25$	3, 4, 5, 6

За ступенем впливу істотні та критичні ризики потребують певних дій, які дадуть можливість уникнути їх. Так як, у нашому випадку відсутні критичні ризики, то аналізуємо лише істотні. Через те, що неприпустимі елементи (за рівнем ризиком) такі ж самі, то одразу все характеризуємо та надаємо необхідні поради.

3. Вибір оптимальної технології – ризик того, що обрана технологія не дозволить отримати необхідний результат. Детальний аналіз існуючих технологій було проведено на етапі дослідження предметної області.

4. Помилки в документації – ризик того, що наявні недопустимі помилки, які не дозволять здати проект. Для того, щоб цього не трапилось необхідна детальна перевірка усієї документації декількома людьми.

5. Помилки проектування – ризик того, що під час проектування обраний помилковий шлях розробки і виконаний зовсім інший проект. Щоб уникнути цього на етапі проектування тісніше співпрацювали із замовником та на певних етапах демонстрували поточні результати.

6. Недостатність кваліфікації для реалізації поставлених задач – ризик того, що проект буде виконано на низькому рівні або не буде виконано вчасно. Уникнути це можна завдяки вивченню інформації на тематичних форумах, відвідуванню конференцій, проходженню онлайн-курсів.

8. Відсутність резервних копій даних – ризик того, що через відсутність резервних копій можна залишитися без проекту. Адже якщо, наприклад, відбудеться збій чи то програмного чи апаратного забезпечення і розробка проекту не збережеться. Тому необхідно як часті створювати резервні копії кожного етапу

Отже, ризики були враховані та здійснено ряд заходів по недопущенню їх впливу на кінцевий результат.

ДОДАТОК Б

В результаті роботи було розроблено програмний додаток визначення медичної дезінформації в онлайн-спільнотах. У цьому додатку знаходиться програмний код.

Таблиця А.1 – Набір тверджень для побудови та навчання нейронної мережі.

rating, "statement"
 false, "Automobile components emit dangerous levels of cancer-causing benzene fumes."
 false, "Hiding finely-ground glass in food is an effective method for killing a person."
 false, "Cancer is caused by a deficiency of vitamin "B17," a condition that can be remedied with nutritional supplements."
 false, "College student who restricts himself to only one foodstuff contracts scurvy."
 false, "Ordinary use of canola oil is dangerous to consumers."
 true, "Some brands of sidewalk chalk have contained high levels of lead."
 false, "vaccines had cancer enzymes in them"
 false, "side effects of the [HPV] vaccine outweigh the positives"
 false, "Each of these ingredients [thimerosal, ovalbumin, formaldehyde, sodium deoxycholate] is a known neurotoxin"
 false, "vaccines cause autism"
 false, "With repeated use of vaccines, cell-mediated immunity weakens"
 false, "If a child has a mitochondrial disorder [...] she is more at risk of having a vaccine pushing her over the edge and causing a bad reaction"
 false, "Measles Vaccines Kill More People than Measles, CDC Data Proves"
 false, "Toxic vaccine ingredients in the MMR vaccine cause encephalopathy"
 false, "Formerly healthy people are suffering serious harm or even dying from taking the flu shot"
 true, "HPV vaccine may lead to elimination of cervical cancer"
 true, "HPV vaccine safe"
 true, "Flu Vaccine effective"
 true, "A virus deadly to dogs has been spreading in the canine population of the U.S."
 false, "Ingesting ""a tablespoon"" of fruit syrup every 15 minutes for an hour can help someone dealing with vomiting or diarrhea."
 false, "Artificially sweetened foods and drinks are less healthy, than their sugary counterparts."
 false, "Turmeric can prevent dementia"
 false, "Diets tailored to one's specific blood type are capable of reducing myriad ailments, improving digestion, enabling weight loss, and providing increased energy."
 false, "Chelsea Clinton's home is actually a medical care facility for Hillary Clinton in order to more effectively hide and treat her health problems."
 false, "The artificial sweetener aspartame is responsible for an epidemic of cancer, brain tumors, and multiple sclerosis."
 false, "Swishing plant oils in your mouth ameliorates a variety of medical ailments."
 false, "Drinking four glasses of water at the beginning of each day will prevent heart attacks and cure various diseases."
 true, "A form of strep infection is capable of affecting the nasal cavity and face."
 true, "lack of sleep increase your chance of developing diabetes"
 true, " vitamin C useful for type-2 diabetes"
 true, "The drug Premarin is produced from the urine of pregnant horses."

Продовження таблиці А.1

false,"Study produced by fast food companies proves the more overweight one is, the healthier one is."

false,"Soup cans are lined with dangerous levels of BPA."

false,"vaccines had cancer enzymes in them"

false,"Cancer Institute Finally Admits Marijuana Kills Cancer"

false,"Marijuana is non-addictive, cures cancer, and grows brain cells"

false,"Recent Study Shows How Sunscreen Causes Cancer, Not the Sun"

false,"Science just confirmed eating turmeric every day could ‘reverse cancer’"

false,"We believe we will offer in a year’s time a complete cure for cancer."

false,"There Are Now 100 Scientific Studies That Prove Cannabis Cures Cancer"

false,"root canals cause cancer"

false," use of mobile phones everyday linked with increased risk of brain cancer"

true,"HPV vaccine may lead to elimination of cervical cancer"

true,"metastasis the dominant reason people die from cancer"

true," moderate alcohol consumption increase risk of cancer"

true," biopsy necessary for prostate cancer detection "

true," water fluoridation safe"

true,"Doctors removed 526 tooth-like structures from a child’s mouth in India."

true," Mexican researcher eliminate HPV in hundreds of women using photodynamic therapy"

true," HPV cause cervical cancer"

true," women get mammograms starting at age 40"

true,"taking low dose aspirin help prevent colorectal cance"

false,"The U.S. Department of Justice admitted that the flu shot is the most dangerous vaccine in America."

false,"A 9-year-old girl was paralyzed as a result of receiving a flu shot."

false,"Tamiflu is responsible for the increased number of deaths this flu season."

false,"Production of the flu shot caused the influenza virus to mutate into a more virulent strain, driving the deadly 2017-2018 flu season."

false,"Peanut oil is secretly present in childhood vaccines and flu shots."

false,"The flu shot is dangerous and full of risks the pharmaceutical companies don't want you to know about."

false,"The CDC and CIDRAP have admitted that Ebola is now airborne."

false,"The U.S. government is planning to implement mandatory Ebola vaccinations for all residents."

true,"A virus deadly to dogs has been spreading in the canine population of the U.S."

true,"Handles of shopping carts are laden with germs."

true,"U.S. Secretary of Defense Donald Rumsfeld owns stock in the company that makes Tamiflu."

true,"Curcuma longa L. (turmeric) of ginger family (Zingiberaceae) has been used also as a safe and active drug for the treatment of diabetes mellitus."

true,"Curcumin possesses multiple therapeutic properties."

true,"Curcumin has antioxidant and cancer preventive properties"

true,"Cinnamon, Cardamom, Saffron, and Ginger had significantly beneficial effects on cholesterol"

true,"Figs are used for gastrointestinal disorders such as loss of appetite, colic, diarrhea, and indigestion."

true,"Figs are used as purgative, aphrodisiac, anti-inflammatory, expectorant, diuretic, and anti-anxiety (mild sedative)."

true,"Cardiovascular disorders are treated with figs"

true,"The figs extracts possess antianaemia and anticarcinogenic activity."

true,"Bauhinia candicans ia a simple tea leaf that could help to balance sugar levels."

true,"Bauhinia candicans can be applied to treat diabetes, hyperglycemia, and polyuria"

Продовження таблиці А.1

true,"Allium cepa (onion) and Allium sativum (garlic) act in an antidiabetic and hypolipidemic manner"
 true,"Dietary consumption of pistachio nuts as a snack has beneficial effects on glycemic control, blood pressure, obesity, and inflammation markers in diabetic patients."

true,"Milk thistle has been used to treat liver and gallbladder diseases. "

true,"the vitamin D3 supplement of 25 µg/day had no beneficial effect on glycaemic indices in healthy overweight or obese women."

true,"bilberries may be valuable as a dietary preventive agent against hypercholesterolemia, probably by virtue of their high anthocyanin content."

true,"High polyphenol chocolate is effective in improving the atherosclerotic cholesterol profile in patients with diabetes"

true,"Increased dietary fish-oil consumption is associated with a reduced risk of coronary heart events "

true,"fish-oil supplement improve renal function without changes in endothelial function, metabolic profiles, blood pressure, inflammation or oxidative stress in patients with type 2 of diabetes mellitus."

true,"Cranberry supplements are effective in reducing atherosclerotic cholesterol profiles"

true,"Cranberry supplements have a neutral effect on glycaemic control in Type 2 diabetic subjects taking oral glucose-lowering agents."

true,"Karela is often used as a treatment for diabetes."

true," karela reduces hyperglycaemia in Type 2 diabetes without stimulation of insulin. "

true," karela consumption can help to control hyperglycaemia"

true,"consumption of powdered cinnamon (Cinnamomum tamala) for a month decreased fasting plasma glucose and improved glucose tolerance in diabetic patients."

true,"Iron is used therapeutically in higher doses to treat or prevent iron deficiency anemia. "

true,"Vitamin C helps your body absorb iron. "

true,"Vitamin B12 helps in cases of vitamin deficient or pernicious anemia. "

true,"Folic acid can be taken for folic acid deficiency, which can cause anemia."

true,"Spirulina, or blue-green algae, may treat some anemias. "

true,"Alfalfa (Medicago sativa), dandelion (Taraxacum officinale) root or leaf, burdock ((Arctium lappa), and yellowdock (Rumex crispus) for mild cases of anemia may help bring levels of hemoglobin into normal range."

true,"Vitamin E, Selenium, Zinc and Quercetin may improve diabetes by returning blood glucose levels to the normal range) and reduce the risk of associated complications"

true,"Biotin (a B-complex vitamin) is helpful for type 2 diabetes"

true,"Vitamin B6 is helpful for both type 1 and type 2 diabetes."

true,"Beta-carotene, Vitamin C, Vitamin E have been shown to improve cholesterol levels in people with type 2 diabetes. "

true," Bitter melon has traditionally been used as a remedy for lowering blood glucose in patients with diabetes."

true,"Bitter melon may decrease serum glucose levels. "

true,"Bitter melon can be extremely dangerous to take when pregnant."

true,"Fenugreek seeds, a spice found in many curry preparations, have been shown to regulate glucose and improve lipid levels in both animals and humans."

true,"cinnamon may play an important role in regulating blood sugar in people with diabetes."

true,"American ginseng appear to lower blood glucose levels"

true,"Getting vaccinated before the flu season starts reduces your chances of getting the flu and helps you recover faster if you do get it."

true,"The best way to protect yourself from the flu is to get an annual vaccine"

true," Annual flu vaccination is recommended if you are 50 or older, have chronic heart, lung, or kidney disease, have a weakened immune system"

true,"For fever reduction and relief of minor aches and pain use ibuprofen (Advil, Motrin) and acetaminophen (Tylenol)"

true,"Elderberry (Sambucus nigra) may help reduce the symptoms of colds and flu. "

Продовження таблиці А.1

```

true,"Eucalyptus is used in many remedies to treat cold and flu symptoms, particularly cough."
true,"Like eucalyptus, peppermint is widely used to treat cold and flu symptoms."
true,"Okra may have some beneficial effect in helping to regulate blood sugar levels."
false,"Okra can “cure diabetes” or eliminate the need for diabetics to take insulin."
true,"Lantus, novolog, and metformin can low blood sugar. "
true,"Metformin (Glumetza, Glucophage, Fortamet) and insulin are used to treat diabetes. "
false," CBD Oil can help control blood sugar for people with diabetes."
false,"guava can help lower diabetes risk"
true,"multivitamin with iron are used to treat anemia"
true,"iron supplements is necessary when anemia"
true,"B12 deficiency causes anemia"
true,"people with pernicious anemia a need B12 injections"
true,"strict vegetarian diet may cause anemia"

```

claims.ipynb

```

import requests
import json
import csv
import pandas as pd
url = 'https://api.monant.fiit.stuba.sk/auth'
payload = "{\n  \"username\": \"parfenenko\",\n  \"password\": \"33ef38b5-181f-4b3f-ba00-
e2479b8bb0db\"\n}"
headers = {
  'Content-Type': 'application/json'}
generatedtoken = requests.request('POST', url, headers = headers, data = payload)
a=generatedtoken.json()
url2 = ' https://api.monant.fiit.stuba.sk/v1/claims?
category=medical&search=cancer&rating=false&page=1'
headers = {
  'Authorization': 'JWT ' + a['access_token'],
  'Content-Type': 'application/json'}
response = requests.request('GET', url2, headers = headers, data = payload,
allow_redirects=False)
todos = json.loads(response.text)
with open('claims1.json', 'a', encoding='utf-8') as f:
    json.dump(todos, f, ensure_ascii=False, indent=4)

```

modul_dezinform.ipynb

```

import numpy as np
import pandas as pd

#завантаження вхідного набору
train_data=pd.read_csv('C:/Users/user/Anaconda3/envs/project_env/input/claims.csv')
test_data = pd.read_csv('C:/Users/user/Anaconda3/envs/project_env/input/posts1.csv')
train_data.head()
print(train_data.shape)

```

```

print(train_data.isnull().sum())
ratings=train_data.rating.unique()
dic={}
for i,rating in enumerate(ratings):
    dic[rating]=i
labels=train_data.rating.apply(lambda x:dic[x])

#розділяємо вхідні дані на тренувальну та тестову вибірки
val_data=train_data.sample(frac=0.1,random_state=121)
train_data=train_data.drop(val_data.index)
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
from keras.utils import to_categorical
texts=train_data.statement

#попередня обробка та токенізація тексту
NUM_WORDS=2000
tokenizer = Tokenizer(num_words=NUM_WORDS,filters='!"#$%&()*+,-./:;<=>?@[\\]^_`{|}~\t\n"',
                      lower=True)
tokenizer.fit_on_texts(texts)
sequences_train = tokenizer.texts_to_sequences(texts)
sequences_valid=tokenizer.texts_to_sequences(val_data.statement)
word_index = tokenizer.word_index
print('Found %s unique tokens.' % len(word_index))

#перетворення слів в числа
MAX_SEQUENCE_LENGTH = 27
X_train = pad_sequences(sequences_train)
X_val = pad_sequences(sequences_valid,maxlen=X_train.shape[1])
y_train = to_categorical(np.asarray(labels[train_data.index]))
y_val = to_categorical(np.asarray(labels[val_data.index]))
print('Shape of X train and X validation tensor:', X_train.shape,X_val.shape)
print('Shape of label train and validation tensor:', y_train.shape,y_val.shape)

#формування числових векторів
from keras.layers import Embedding
EMBEDDING_DIM=200
vocabulary_size=min(len(word_index)+1,NUM_WORDS)
embedding_layer = Embedding(vocabulary_size,
                            EMBEDDING_DIM)

#створення нейронної мережі
from keras.layers import Embedding
from keras.models import Model
from keras.layers import Input, Dense, Embedding, Conv1D, MaxPooling1D, Dropout,concatenate
from keras.layers.core import Reshape, Flatten
from keras.callbacks import EarlyStopping
from keras.optimizers import Adam
from keras import regularizers
from keras.regularizers import l2
from keras.layers.normalization import BatchNormalization

```

```

sequence_length = X_train.shape[1]

inputs = Input(shape=(sequence_length,))
embedding = embedding_layer(inputs)
inp_norm = BatchNormalization(axis=1)(embedding)

#параметри
batch_size=128
num_epochs = 20
l2_lambda = 0.0001
filters=32
pool_size =2
kernel_size=3
ens_models = 2
num_classes=2

convs = []
for i in range(ens_models):
    l_conv1 = Conv1D(filters,kernel_size, strides=1,
padding='valid',data_format='channels_first',init='he_uniform',W_regularizer=l2(l2_lambda),activation='relu')(inp_norm)
    l_conv1 = BatchNormalization(axis=1)(l_conv1)
    #l_conv2 =
Conv1D(filters,kernel_size,border_mode='same',data_format='channels_first',init='he_uniform',W_regularizer=l2(l2_lambda),activation='relu')(l_conv1)
    #l_conv2 = BatchNormalization(axis=1)(l_conv2)
    l_pool = MaxPooling1D(pool_size)(l_conv1)
    convs.append(l_pool)
l_merge = concatenate(convs, axis=1)

x = Dropout(0.1)(l_merge)
x = Flatten()(x)
x = Dense(128, activation='relu')(x)
x = BatchNormalization(axis=1)(x)
x = Dropout(0.2)(x)
output = Dense(num_classes,init='glorot_uniform',W_regularizer=l2(l2_lambda),
activation='softmax')(x)
# this creates a model that includes
model = Model(inputs, output)

#компіляція моделі
model.compile(loss='categorical_crossentropy',
optimizer='adam',
metrics=['acc'])

#навчання моделі
model.fit(X_train, y_train, batch_size=batch_size, epochs=num_epochs, verbose=1,
validation_data=(X_val, y_val))

#визначення точності навчання
test_loss, test_acc = model.evaluate(X_val, y_val)

```

```

print( "Test accuracy:", test_acc)
#використання навченої моделі із тестовими даними
sequences_test=tokenizer.texts_to_sequences(test_data.Content)
X_test = pad_sequences(sequences_test,maxlen=X_train.shape[1])
y_pred=model.predict(X_test)
print(y_pred)
np.argmax(y_pred[0])

posts_get.ipynb
import os
import csv
import json
import requests
from urllib.request import urlopen
#аутентифікація
headers = {'Authorization': 'Basic !PFPswZ0Fdj3YucedJ',
           'Username': 'nastya_blog',
           'Password': '!PFPswZ0Fdj3YucedJ'
          }
#маршрут
url_link=http://medonline.zzz.com.ua/wp-json/wp/v2/posts?status=pending
#запит на отримання постів
response = requests.request( 'GET',url_link, headers=headers, allow_redirects=False)
with urlopen(url_link) as url:
    data = url.read()
print (data)

#запис до файлу
filename = "posts json1.txt"
file_ = open(filename, 'wb')
file_.write(data)
file_.close()
def save_to_file (fn, row, fieldnames):
    if (os.path.isfile(fn)):
        m="a"
    else:
        m="w"
    #конвертація в csv файл
    with open(fn, m, encoding="utf8", newline=" ") as csvfile:
        writer = csv.DictWriter(csvfile, fieldnames=fieldnames)
        if (m=="w"):
            writer.writeheader()
            writer.writerow(row)
with open(filename) as json_file:
    json_data = json.load(json_file)
#збереження csv файлу
for n in json_data:
    r={}
    r["Title"] = n['title']['rendered']
    r["Content"] = n['content']['rendered']
    save_to_file ("posts.csv", r, ['Title', 'Content'])

```

```
into_post.ipynb
import MySQLdb
conn = MySQLdb.connect('localhost', 'sitediplom', 'Nana20021997', 'wp_posts')
cursor = conn.cursor()
cursor.execute("
LOAD DATA INFILE 'C:/Users/user/Anaconda3/envs/project_env/input/posts1.csv'
INTO TABLE wp_posts
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 ROWS;")
conn.close()
```