# РЕФЕРАТ

Пояснювальна записка містить 73 сторінки, 24 рисунка, 7 таблиць, 4 додатка, 26 посилань.

Метою дослідження є полегшення виявлення елементів дезінформації за рахунок створення методу та алгоритму для перевірки потоку текстових даних на наявність елементів дезінформації у вигляді лінгвістичних конструкцій та оборотів, які вказують на неправдивість представленої інформації.

Для досягнення поставленої мети необхідно виконати наступні завдання:

- виконати аналіз існуючих алгоритмів та методів комп'ютерної лінгвістики та машинного навчання для класифікації текстових потоків даних та виявлення елементів дезінформації;
- розробити алгоритм первинної обробки тексту для збільшення точності визначення елементів дезінформації;
- розробити метод виявлення елементів дезінформації в текстових потоках даних;
- виконати програмну реалізацію розробленого методу виявлення елементів дезінформації в текстових потоках даних;
- провести аналіз отриманих результатів для оцінки якості;
- провести дослідження ефективності алгоритму.

Актуальність. З розповсюдженням Інтернету та соціальних медіа зараз доступна кількість новин, статей та іншого тексту онлайн. Цей величезний обсяг інформації постав під загрозу правдивість новин.

Підроблені новини - це будь-яка форма помилкової інформації чи контенту, що поширюється в мережі Інтернет, для впливу на погляд людей на певну подію чи інформацію. Виявлення фальшивих новин у цифровому світі є важливим завданням у подоланні широкого розповсюдження чуток та упереджень. Багато досліджень було проведено для виявлення елементів дезінформації для англійської мови, проте українська та російська мови не має досліджень у цій галузі. Такі компанії, як Facebook, Twitter та Google, стикаються з проблемою вирішення цієї проблеми, щоб забезпечити

платформу, де люди можна довіряти вмісту стрічки новин. Вплив фейкових новин було таким глибоко вкорінене в суспільстві, що це навіть вплинуло на вибори в США 2016 року. Також багато неправдивої інформації поширюється протягом війни України в зоні АТО, що призводить до дестабілізації населення, поширення неправильних думок, відображення фейкової картини перебігу подій.

Отже, необхідною задачею є створення інструменту перевірки текстової інформації на наявність елементів дезінформації для інформаційної безпеки та аналізу новин, які поширюються для дестабілізації та обману населення.

Предметом дослідження є методи виявлення елементів дезінформації в текстових потоках даних.

Методами дослідження є методи комп'ютерної лінгвістики та машинного навчання для виявлення елементів дезінформації.

Науковою новизною є розробка методу виявлення елементів дезінформації в потоках даних з підтримкою обробки текстів української та російської мови.

Основні положення роботи доповідались і обговорювались на IX науково-технічній конференції «Інформаційні моделі, системи та технології»

Ключові слова: ДЕЗІНФОРМАЦІЯ, ОБ'ЄМНІ ТЕКСТОВІ ДАНІ, КЛАСИФІКАЦІЯ, МАШИННЕ НАВЧАННЯ.

# ABSTRACT

The explanatory note contains 73 pages, 24 figures, 7 tables, 4 appendices, 26 links.

The purpose of the study is to facilitate the detection of elements of misinformation by creating a method and algorithm to check the flow of textual data for the presence of elements of misinformation in the form of linguistic constructions and turns that indicate the falsity of the information.

To achieve this goal it is necessary to perform the following tasks:

- perform an analysis of existing algorithms and methods of computational linguistics and machine learning to classify textual data flows and identify elements of misinformation;
- develop an algorithm for primary text processing to increase the accuracy of determining the elements of misinformation;
- develop a method for detecting elements of misinformation in text data streams;
- perform software implementation of the developed method of detecting elements of misinformation in text data streams;
- analyze the results obtained to assess quality;
- conduct research on the effectiveness of the algorithm.

Relevance. With the spread of the Internet and social media, a number of news, articles and other text are now available online. This vast amount of information has jeopardized the veracity of the news.

Fake news is any form of false information or content that is distributed on the Internet to influence people's views on an event or information. Detecting fake news in the digital world is an important task in overcoming rumors and prejudices. Many studies have been conducted to identify elements of misinformation for English, but Ukrainian and Russian have no research in this area. Companies like Facebook, Twitter and Google are facing a challenge to provide a platform where people can trust the content of the news feed. The influence of fake news was so deeply

ingrained in society that it even affected the 2016 US election. Also, a lot of false information is spread during the war in Ukraine in the ATO zone, which leads to destabilization of the population, the spread of misconceptions, reflecting a fake picture of events.

Therefore, a necessary task is to create a tool to check textual information for the presence of elements of misinformation for information security and analysis of news that are distributed to destabilize and deceive the population.

The subject of research is methods of detecting elements of misinformation in text data streams.

The research methods are computer linguistics and machine learning methods to identify elements of misinformation.

A scientific novelty is the development of a method for detecting elements of misinformation in data streams with support for processing Ukrainian and Russian texts.

The main provisions of the work were reported and discussed at the IX scientific and technical conference "Information models, systems and technologies"

Keywords: DISINFORMATION, LARGE TEXT DATA, CLASSIFICATION, MACHINE LEARNING.

ЗМІСТ

# ВСТУП

Популяризація Інтернет спричинила безпрецедентне розповсюдження інформації. Щоденно близько 2500 петабайт інформації генерується на просторах світової мережі [1]. Суттєва частина цього величезного обсягу даних є дезінформацією, що ставить під загрозу правдивість новин.

Дезінформація – це тип інформації, який створюється і поширюється з наміром введення кінцевого користувача в оману стосовно реального стану справ. Постійне споживання дезінформації призводить до викривленої реальності, через це поширення неправдивих новин зазвичай відбувається у пропагандистських, військових або комерційних цілях.

На сьогодні яскраві приклади дезінформації щоденно зустрічаються в соціальних мережах. Цифрові гіганти Meta, Google, Twitter володіють платформами де щоденно поширюються мільярди новин, тому повинен існувати механізм що забезпечує їх достовірність.

Актуальним прикладом дезінформації є первинні новини після спалаху Коронавірусу (COVID-19). В мережі Інтернет почали генеруватися різноманітні, безпідставні теорії щодо походження хвороби. За відсутності досліджень люди спекулювали щодо різноманітних аспектів захворювання. Деякі стверджували, що вірус є біологічною зброєю, а результуюча потреба в вакцинації – схема контролю населення.

Завданням цієї роботи полягає у створення системи що забезпечує перевірку текстової інформації, виявляє елементи дезінформації, аналізує шляхи її розповсюдження і, як результат, надає інформаційну безпеку кінцевому користувачу.

# 1 ОБЛАСТЬ ЗАСТОСУВАННЯ ТА ЗАГАЛЬНИЙ ОГЛЯД МЕТОДІВ ТА ПІДХОДІВ ДО ВИЯВЛЕННЯ ЕЛЕМЕНТІВ ДЕЗІНФОРМАЦІЇ

## 1.1 Загальні відомості про дезінформацію та аналіз літератури

Сьогодні люди часто вживають термін «фейкові новини» при першій можливості, але що ж це таке насправді? За даними Columbia Journalism Review, існує шість типів фейкових новин [2]. Підроблений вміст – це підроблені фотографії або відео, які створюють фейковий сценарій, але поширюються як реальність. Неправдива інформація – візуальний пост або повідомлення з неправильною інформацією що, як правило, поширюється без перевірки. Автентичний матеріал, використаний у неправильному контексті – зображення або відео іншої події яка сталася в минулому, як-от арешт чи мітинг, яке презентується ніби щось актуальне, вводячи глядача в оману. Іншою формою фейкових новин є сайти самозваних новин, які видають себе за справжні сайти.

Це означає, що фейкові новини — це не лише один тип контенту, що надходить з одного джерела, а різні типи підроблених засобів масової інформації, часто сенсаційних, створених з метою викривлення реальності. глядачі бачать цю фальшиву реальність і діляться нею, увічнюючи легку вірусність фейкових новин.

Проблема в тому, що люди думають, що зможуть розпізнати дезінформацію, коли вона з’явиться на їхньому екрані. Звіт за 2019 рік, опублікований Pew Research, показав що хоча респонденти були впевнені, що зможуть розпізнати фейкові новини, побачивши їх, близько 60% зізналися що поширювали дезінформацію навіть не знаючи, що це були фейкові новини [3]. Це означає, що зловмисники, які створюють фейковий контент, контент в неправильному контексті і дезінформацію, легко презентують брехню, в яку вони хочуть, щоб люди вірили.

Фальшиві новини створюють фальшиву реальність, на яку купуються не лише один чи двоє людей, а тисячі, реальність якою ті, хто створює фейкові

новини, можуть маніпулювати як їм заманеться. Якщо пустити все на самотік, то ця брехня може загрожувати стабільності суспільства, або в екстремальних випадках стати смертельною.

Раніше згаданий звіт  Pew Research показав, що люди оцінюють проблему вигаданих новин та дезінформації вище, ніж проблеми насильницьких злочинів, зміни клімату чи расизму. Це проблема настільки великого масштабу, що половина опитаних людей вважає що поширення фальшивих новин повинно вважатися кримінальним злочином.

Рішення проблеми дезінформації починається з більшої прозорості щодо автора та організації, що стоїть за новиною. Згідно з опитуванням журналу Forbes, 44% респондентів заявили, що їхня довіра до вмісту збільшилася б, якби вони знали, хто його створив і звідки він береться, а 64% хотіли б отримати звіт про те, як він змінився з часом [4].

На даний момент в Інтернеті — як це і було з самого початку — немає системи прозорості та звітування. Тому дуже важливо, щоб люди почали перевіряти факти, що зустрічаються їх у новинних стрічках, це чудова практика щоб стати більш поінформованим користувачем ЗМІ. Пошук автора та організації ким була написана стаття, також може допомогти підвищити довіру. Існують також технологічні компанії, які створюють рішення, які мають на меті допомогти «заохочувати медіа грамотність» і розкривати фейкові новини.

Інші способи підвищити довіру включають надання читачам доступу до історії версій вмісту, щоб показати, як він змінився з моменту його створення, і можливість легко дізнатися більше про автора. Поширеною також є думка, що пошукові системи можуть вжити заходів для зменшення вмісту без автора або не пов’язаного з організацією. Ідея в не в тому, щоб заглушити голоси, а в тому, щоб заохочувати більше прозорості щодо того, хто ці голоси та чи правда те, що вони говорять.

Отож хто несе відповідальність за втручання та вирішення проблеми фейкових новин? Саме це питання мотивує створення системи з виявлення елементів дезінформації. Сучасні дослідження фальшивих новин є обмеженими, а саме визначення фейкових новин є нечітким, тому варто

розглянути існуючі методи виявлення дезінформації, провести глибокий аналіз та дослідити правила за якими підроблені новини створюються та поширюються.

## 1.2 Загальний огляд методів комп'ютерної лінгвістики для виявлення елементів дезінформації

Перший крок з вирішення поставленої задачі полягає у виявленні принципів поширення дезінформації. Загальновідомими є низка досліджень [5], що виявили певні правила за якими конструюється фальшива новина. Фейковий контент здебільшого подається в урізаному вигляді, інформація є об'єктивно неповною, текст містить багато скорочень, а тон тексту, як правило, атакує певну сутність. Фейкові статті зазвичай базуються на емоційному посилі та ігнорують сухі факти, якщо вони суперечать негативному характеру атакованої сутності.

Надалі приведені деякі актуальні дослідження з ідентифікації елементів дезінформації, що використовують методи машинного навчання та комп'ютерної лінгвістики.

### 1.2.1 Проблема виявлення спаму

Багато дослідників за останнє десятиліття зосередили свої зусилля на вирішення проблеми фільтрації спаму в соціальній мережі Twitter, оскільки саме ця платформа пропонує найширший інструментарій для роботи з великими масивами даних.

Прикладом спроби вирішення проблеми спаму є класифікатор на основі вмісту твітів що був побудований відповідно до лінгвістичного аналізу

повідомлень [6]. Але він не міг генерувати набір порівнянних результатів, оскільки в його механізмі використовувався лише один алгоритм. Останнім часом більшість досліджень акцентують увагу на створенні бінарних класифікаторів на основі машинного навчання з введенням статичних ознак.

Унікальні ознаки повідомлень можна отримати з API потокової передачі Twitter і обчислити за допомогою об’єкта JSON, і вони включають атрибути на рівні облікового запису (кількість підписок, кількість підписників і вік облікового запису) та атрибути на рівні користувача (наприклад, кількість URL-адрес, цифри, хештеги у твіті відповідно). Однак існували деякі проблеми з вилученням ознак і незадоволеною точністю. Під час процедури збору даних було помічено, що спам у Twitter буде дрейфувати, а ознаки можна легко сфабрикувати. Крім того, підсумовуючи результати існуючих досліджень, середнє значення точності виявлення спаму досягає лише 85% або близько того.

Інша методика охоплює служби чорного списку, але було показано, що понад 90% користувачів можуть натискати шкідливі URL-адреси до того, як їх занесли в чорний список. У той же час методи внесення в чорний список займають надзвичайно багато часу через участь окремих осіб у розпізнаванні небажаної інформації [7].


Рисунок 1.1 – Вектори виявлення спаму на платформі Twitter

Щоб впоратися з проблемами єдиного підтримуючого алгоритму, вилучення ознак, нестачі точності та низької швидкості, було запропоновано метод класифікації на основі глибокого навчання.

Цей метод складається з кількох етапів. По-перше, застосовується Word2Vec для попередньої обробки твітів замість функції вилучення, де прийнята техніка є передовим методом обробки мови в глибокому навчанні і може конвертувати слово або документ у репрезентативний вектор. Після цього, на основі кількох алгоритмів машинного навчання будується двійкова модель що розрізняє «спам» і «не спам». На фінальному етапі призначається налаштування параметрів фільтрації спаму.

На сьогодні вищевказаний метод що використовує глибинне навчання активно використовується для виявлення спаму на платформі Twitter. Цей метод підвищує точність та швидкість виявлення спаму і також вирішує проблему вилучення унікальних ознак.

### 1.2.2 Синтаксичний аналіз

Одним з традиційних методів виявлення спаму є аналіз граматики мови та синтаксису [8]. Ймовірнісна безконтекстна граматика (PCFG) – це метод що розділяє структури в тексті визначені граматикою. Метод трансформує речення у структури даних, зазвичай дерева, що описують граматично синтаксичну одиницю. На сьогодні даний метод найчастіше використовується для аналізу емоційного посилу речення.

### 1.2.3 Семантичний аналіз

Дуже часто правдивість новини чи будь-якого тексту можна передбачити порівнявши їх з подібними текстами, або ж проаналізувавши коментарі. Якщо подібні тексти вагомо відрізняються, або ж суперечать поточному, то такі тексти з великою долею ймовірності є фальшивими або спотвореними. Відповідно коментарі аналізуються подібним методом.

Даний метод не є широко використовуваним, оскільки автоматизація пошуку подібних текстів є практично неможливою. Також проблема ускладнюється тим, що певні однакові слова в різних контекстах означають різні речі.

### 1.2.4 Використання метаданих

Якщо ви коли-небудь реагували на публікацію у Facebook, ретвітили в Twitter або коментували історію в Instagram, то ви не тільки успішно використовували ці комунікаційні інфраструктури, але й створили власний цифровий слід «метаданих», який також відомий як «індекс людської поведінки».

Наша діяльність на соціальних платформах – наприклад, вибране, лайки, ретвіти, коментарі та реакції використовуються в основному для реклами, але є темна сторона, де маніпулятори, боти та люди, які займаються дезінформацією та дезінформацією, намагаються видати себе за реальних людей, щоб обманути алгоритми що підтримують порядок в соціальних мережах.

Дослідники в сфері маніпулювання медіа визначають це сукупністю практик, які створюють, покладаються на чи навіть грають із поширенням даних у соціальних мережах, використовуючи нові обчислювальні та алгоритмічні механізми організації та класифікації.

Літом 2018 року New Knowledge [9], місцева компанія з дослідження даних в Остіні, штат Техас, опублікувала висновки про пов’язаних з Росією Twitter-ботів і фальшивих акаунтів у Facebook, які використовувалися для

маніпулювання публічним дискурсом під час виборів у США 2016 року. Наявність неправдивих даних про діяльність мала ряд непередбачуваних наслідків для користувачів соціальних мереж і суспільства, від фальсифікації виборів до зміни політичного дискурсу навколо соціальних дебатів, таких як імміграційна реформа та расова політика.

Обробка даних стає все більш шкідливою, тому що більше половини населення отримує свої новини в основному з соціальних мереж. І ця проблема не зникне найближчим часом. Насправді, проблема тільки погіршиться, оскільки боти починають імітувати людську активність у соціальних мережах, щоб виглядати людяніше. Вони стають все більш і більш складними і подекуди можуть створювати дані, що виглядають як повністю залучена до дискусії людина.

Оскільки ті, хто займається обробкою даних, стають розумнішими і хитрішими у імітації того, що виглядає як справжня поведінка в соціальних мережах, що це означає для кінцевого користувача? У майбутньому користувачам стане набагато важче відрізнити «справжніх» користувачів і автентичну діяльність облікового запису від підробок, спаму та шкідливих маніпуляцій.

Метадані є одним з механізмів покликаних для запобігання цього майбутнього. Функції метаданих, такі як назва облікового запису, зображення облікового запису, кількість лайків, теги та дата публікації, можуть надавати сигнали частоти активності, які служать підказками щодо підозрілої активності. Читання метаданих, що оточують облікові записи ботів часто виявляє наміри, промахи та шум, що може додатково виявити автоматизовані маніпуляції.

Визначаючи та розуміючи тактику дезінформації як обробку даних, дослідники інформації можуть читати метадані соціальних мереж так само уважно, як і алгоритми, і, можливо, з більшою точністю, ніж інструменти модерації платформ. Уважно вивчаючи метадані, такі як швидкість облікового запису, активність, кількість підписників/аудиторій, позначки часу публікації, медіа-контент, біографії користувачів і дані про місцезнаходження призвели до

того, що дослідники соціальних мереж і технічні журналісти виявили тисячі фальшивих акаунтів.

Тому, хоча створенню нових ботів і фейкових облікових записів запобігти фактично неможливо, виявлення і ідентифікація дезінформації з цілком реальною задачею. Досліджуючи низку прикладів дезінформації, коли метадані платформи обробляються, видаляються, експлуатуються або збовтуються, ми додамо до зростаючого набору цифрових методів для виявлення кампаній дезінформації та маніпуляцій, а також роботи з даними, керовану ботами, тролями, sockpuppets, і посилення висвітлення гіперпартійних новин на платформах. На сьогодні існує велика кількість програмних рішень для збору та аналізу метаданих.

Alation пропонує платформу для широкого спектру рішень з аналізу даних, включаючи пошук і виявлення даних, управління даними, управління даними, аналітику та цифрову трансформацію. Продукт має механізм поведінкового аналізу, вбудовані можливості співпраці та відкриті інтерфейси. Alation також профілює дані та контролює використання, щоб гарантувати, що користувачі мають точне уявлення про точність даних. Платформа також надає уявлення про те, як користувачі створюють і обмінюються інформацією з необроблених даних.

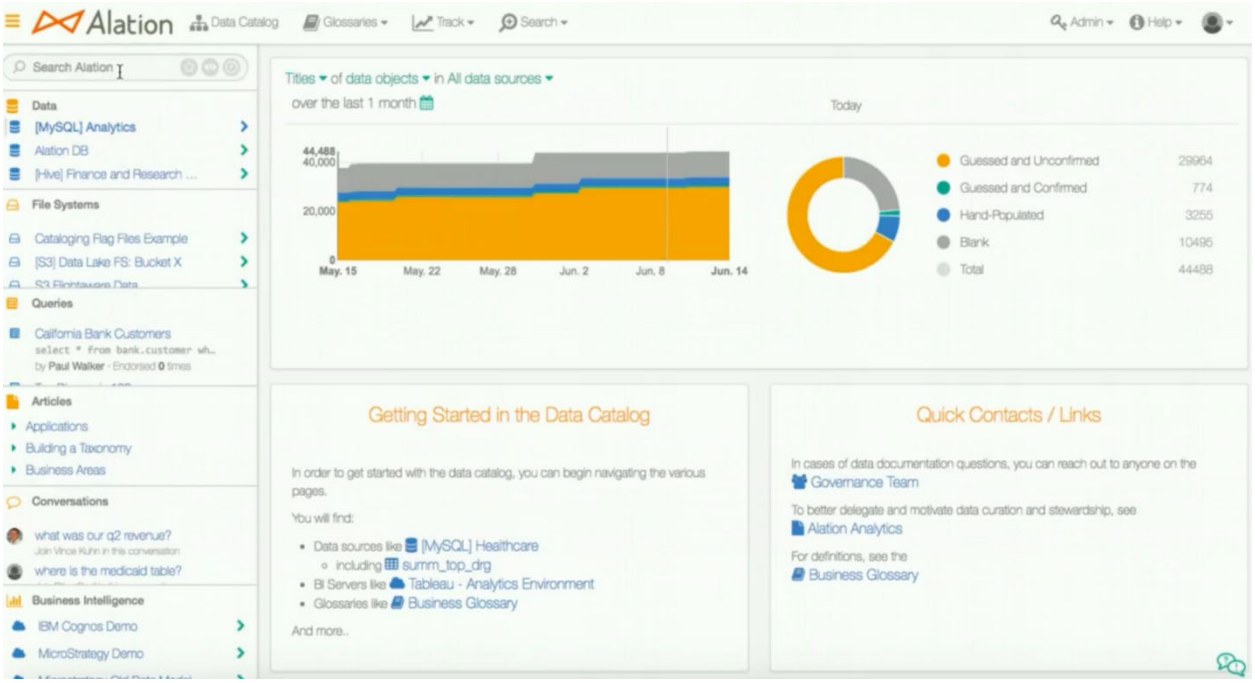


Рисунок 1.2 – Веб-інтерфейс платформи Alation

InfoSphere Information Server IBM сховище метаданих, яке зберігає метадані з зовнішніх інструментів і баз даних дозволяючи обмін даних між ними. Користувачі можуть імпортувати метадані в репозиторій з кількох джерел, експортувати метадані різними методами та передавати метадані між проектними, тестовими та виробничими репозиторіями. Зміни, які вносяться в репозиторій, автоматично вносяться по всьому набору і використовує стандартну технологію реляційної бази даних.

InfoSphere Information Server надає можливості масової паралельної обробки (MPP) для високо масштабованої та гнучкої інтеграційної платформи, яка обробляє всі обсяги даних, великі та малі.

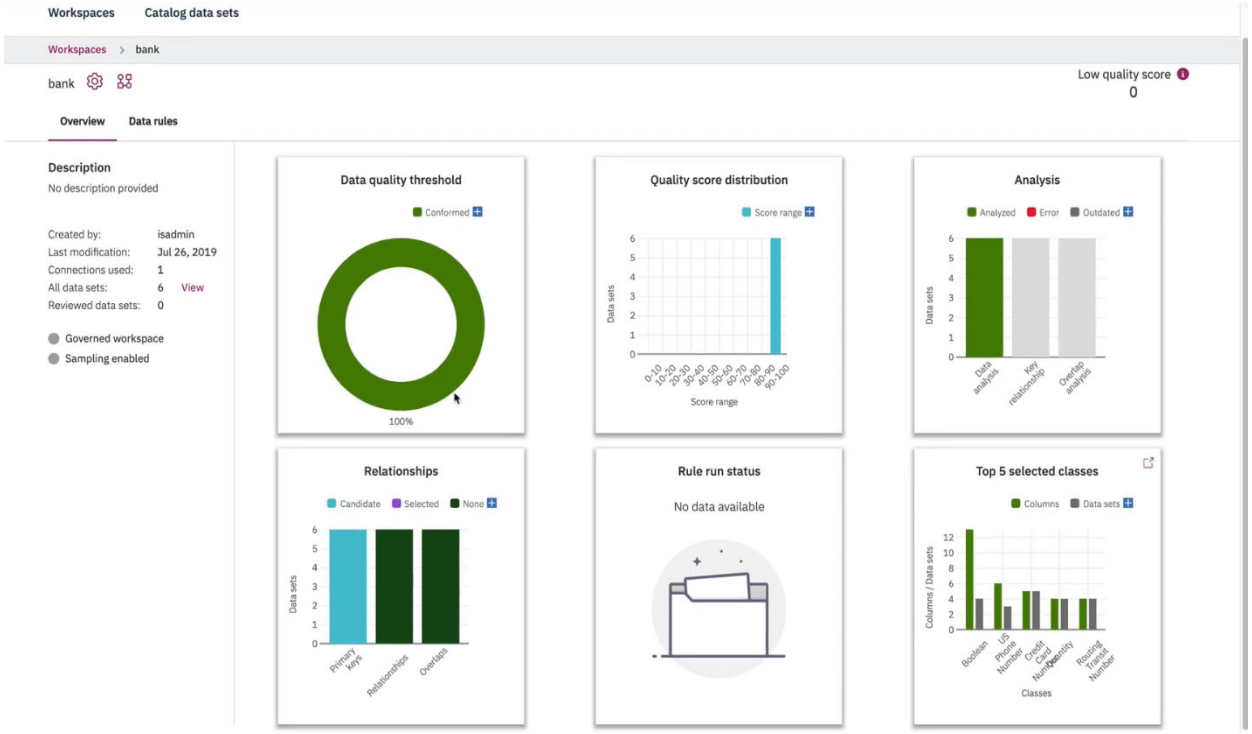


Рисунок 1.3 – Веб-інтерфейс  InfoSphere Information Server

Oracle Enterprise Metadata Management — це платформа керування метаданими, яка може збирати та каталогізувати метадані будь-якого постачальника. Продукт дозволяє здійснювати інтерактивний пошук і перегляд метаданих, а також надає походження даних, аналіз впливу, семантичне визначення та аналіз семантичного використання для будь-якого активу метаданих у каталозі.

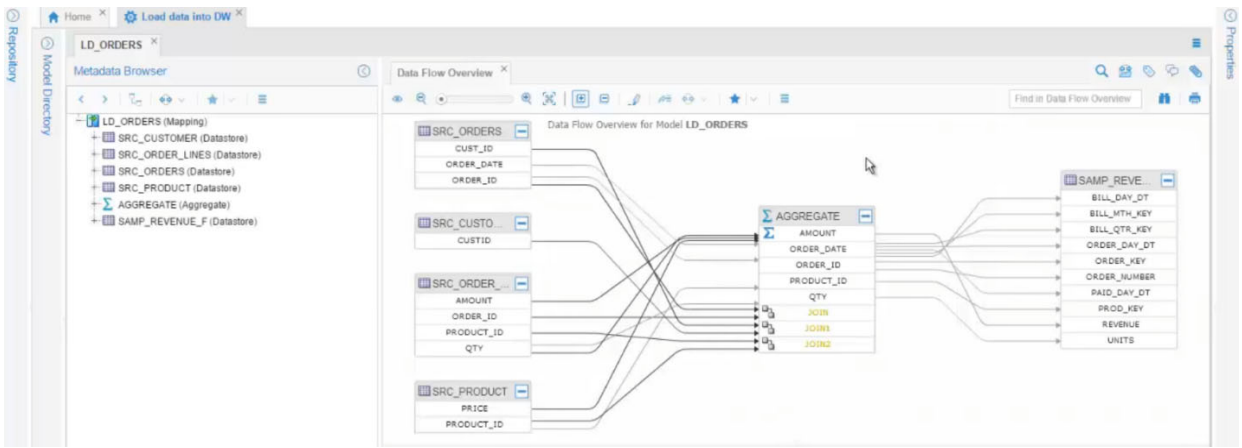


Рисунок 1.4 – Веб-інтерфейс Oracle Enterprise Metadata Management

### 1.2.5 Ручна перевірка фактів

Ручна перевірка неправдивих заяв, чуток і фейкових новин в Інтернеті відіграє важливу роль у стримуванні їх поширення. Можна виділити два широкі класи ручної перевірки: використання веб-сайтів для перевірки фактів і виконання ручної перевірки на конкретних сайтах соціальних мереж.

Веб-сайти перевірки фактів (наприклад, Snopes, Politifact, Emergent) забезпечують верифікацію заяв які вони знайшли або отримали від користувачів. Вони мають перевагу залучення кваліфікованих журналістів та інших професіоналів, які можуть досліджувати та перевіряти заяви та новини. Однак у них є деякі мінуси.

По-перше, як і в освіті, процес переклав відповідальність на особистість. Дослідження [10] вказують на те, що люди навряд чи перевірять історію, яка відповідає їхнім попереднім переконанням. Перевірка фактів може бути навіть контрпродуктивною, оскільки перевірка фактів з історією чи чутками призводить до знайомства з нею, а знайомство породжує не зневагу, а прийняття. Загальноприйнятою є думка що не варто повторювати міф чи чутку, якщо її потрібно спростувати. Скоріше слід повідомляти правильні факти, не згадуючи неправдиву інформацію.

Великі технологічні компанії та сайти соціальних мереж відповіли на соціальний тиск і поширене переконання, що вони зіграли певну роль у сприянні чи, принаймні, не утриманні поширення фейкових новин, оголосивши, що вони наймуть більше модераторів контенту. Ручна перевірка є бажаною оскільки, вона гарантує точну перевірку заяв. Однак вона має багато потенційних підводних каменів, починаючи від можливості поширення упередженості модераторів і закінчуючи психічними збитками, накладеними на осіб, які здійснюють перевірку. Facebook співпрацює з організаціями, що перевіряють факти, щоб зменшити та стримати вплив фейкових новин. У звіті

про співпрацю 2018 року [11] йдеться про відсутність повноцінного успіху, що випливає з відсутності спільних цілей. Організації партнери були стурбовані тим, що ці зусилля не були прозорими ні для них, ні для широкої громади. Також було виявлено і задокументовано проблеми з організаціями що перевіряють факти, зокрема: відсутність координації між собою; надмірна залежність від людського досвіду без, в деяких випадках, плану довгострокової стійкості; або відсутність заходів запобігання дезінформації.

### 1.2.6 Автоматична перевірка фактів

Автоматична перевірка має очевидні переваги: її можна зробити в набагато більшому масштабі, і вона позбавляє модераторів від необхідності сортувати в кращому випадку неприємний вміст. Ця форма перевірки стосується вмісту та тверджень у самій історії, а не метаданих, таких як джерело чи швидкість поширення.

Обчислювальна перевірка фактів намагається знайти неперевірені твердження в історії чи чутках і звірити їх з надійними джерелами. Дослідники Ciampaglia та ін. [12] знаходять фактичну інформацію, перетворюючи Вікіпедію на мережу графіків знань. Неперевірені твердження можна перевірити в цій мережі. Твердження, яке відомо у Вікіпедії як істинне, буде представлено як край графа знань або матиме його суб’єкт і об’єкт, пов’язані через короткий шлях у графі. Імовірно, неправдиві твердження не повинні знаходитися як зв’язані на графіку.

Дослідники Jaradat та ін. [13] створили ClaimRank, обчислювальну систему, яка виявляє заяви, які можуть потребувати перевірки (доступна як для арабської, так і для англійської мов). Твердження, або новини можна надіслати на веб-сайти перевірки фактів (які зазвичай використовують людей для перевірки) або в автоматичні системи. Одна з таких систем знаходить документи, які можуть мати відношення до даного твердження, та фрагменти

доказів. Хоча система не є повністю автоматичною, вона може значно спростити роботу модераторів.

Інша форма автоматичної перевірки передбачає оцінку мови самої історії, тобто пошук у мові оповідання сигналів, які вказують на перебільшені твердження, надмірно емоційну мову або стиль, який не зустрічається в основних джерелах новин. Це, по суті, проблема класифікації тексту, яку зазвичай вирішують лінгвісти-комп'ютери, використовуючи засоби НЛП. Дослідники Potthast та ін. [14] описують цей тип класифікації як визначення фейкових новин на основі стилю, на відміну від контекстно-орієнтованого або факт-орієнтованого виявлення.

### 1.2.7 Використання НЛП

Інтуїтивним підходом до проблеми фейкових новин у НЛП була б класифікація тексту новин на підроблені та легітимні. Це особливо стосується випадку повного тексту – на відміну від твітів чи заголовків, що розповсюджуються в соціальних мережах, – тому що класифікація тексту ґрунтується переважно на мовних характеристиках довшого тексту. Виявлення обману в тексті має довгу історію з НЛП, і фейкові статті новин можна вважати категорією оманливого тексту [15].

Методи, що використовуються для класифікації тексту, варіюються від класичних алгоритмів машинного навчання, що використовують набір попередньо визначених лінгвістичних функцій, до сучасних моделей нейронних мереж, які в основному покладаються на попередньо навчені вектори слів і вбудовані уявлення, що виникають у результаті обробки великої кількості текстових даних.

У НЛП підхід, заснований на чітких ознаках, який передбачає виділення та аналіз мовних сигналів для ідентифікації конкретних цільових явищ (наприклад, різниця між підробленим та справжнім відгуком про товар), був

дуже потужною моделлю з результатами, які можна відносно інтерпретувати. Такі функції, як n-грами, маркери суб'єктивності та полярності, лексико-семантичні класи, синтаксичні або дискурсивні функції, досліджувалися в багатьох роботах щодо виявлення обману і класифікації новин [16][17]. Ці функції можна використовувати з різними традиційними контрольованими алгоритмами. Моделювання на основі ознак зазвичай включає розробку ознак і етап вибору ознак.

На основі порівняльних експериментів у різних програмах машинного навчання також було показано, що продуктивність цих класичних моделей у певний момент стає плато зі збільшення розміру навчальних даних. Таким чином, у проблемах, де доступні великі дані, перевага надається моделям глибоких нейронних мереж, оскільки вони зазвичай досягають вражаюче кращих результатів.

### 1.2.8 Моделі глибокого навчання

В областях де доступні широкомасштабні навчальні дані, глибоке навчання взяло на себе більшість завдань НЛП. У класифікації тексту рекурентні нейронні мережі (RNN), згорткові нейронні мережі (CNN) та моделі Attention конкурують з моделями на основі ознак. RNN здатні кодувати послідовну інформацію і найбільше підходять для моделювання семантики короткого тексту. CNN складаються з шарів згортки та об’єднання, які забезпечують абстракцію вхідних даних.

Ці моделі використовуються в конкретних завданнях НЛП, де наявність або відсутність ознак є більш важливим фактором, ніж їх розташування або порядок. Наприклад, наявність конкретних слів і фраз в огляді продукту зазвичай свідчить про позитивний чи негативний відгук. Таким чином, CNN добре підходять для класифікації довшого тексту. Моделі нейронних мереж

також застосовувалися в попередніх роботах у сфері дезінформації та фейкових новин [18].

Усі провідні методи машинного навчання для класифікації тексту, включаючи моделі на основі функцій та нейронних мереж, значною мірою керуються даними. Тому навчальні дані є першою вимогою для побудови цих моделей. Якісні навчальні дані для виявлення дезінформації мають складатися з збалансованого, достатньо різноманітного та ретельно позначеного набору легітимних та фейкових новинних статей.

Хоча створення такого набору даних може здатися тривіальним, у наступному розділі пояснюються проблеми зі збором такого набору даних. Початкові експерименти показують, що наявних даних все ще недостатньо для створення надійної системи виявлення дезінформації.

## 1.3 Пошук тренувальних даних для нейронних мереж

Перше питання, на яке ми маємо відповісти, вирішуючи питання виявлення фейкових новин за допомогою класифікації текстів, — це те, що саме ми розглядаємо як репрезентативний екземпляр фейкових новин. В інших областях, пов’язаних з оманливим текстом, наприклад виявлення підробленого відгуку про продукт, при позначенні підроблених випадків можуть бути розроблені об’єктивні критерії: відгук, написаний кимось, хто не купував або не використовував продукт, або кимось, кого продавець завербував для конкретного продукту.

Підроблені новини також можна визначити як статті новин, написані дилетантами (а не журналістами), завербованими з прямою метою створення вмісту на користь чи проти організації чи політики, для просування певної ідеї або для фінансової вигоди, наприклад, для залучення кліків для реклами. Професійні журналісти також можуть вигадувати історії з різних причин. Одним з прикладів таких випадків є Клаас Релоціус, журналіст Der Spiegel у Німеччині, який, як було встановлено, вигадував історії, подробиці та цитати з багатьох джерел протягом тривалого періоду часу [19]. У цьому контексті автори та їхні наміри розглядаються як ключовий фактор для визначення чи є стаття підробкою. У даному дослідженні акцент зроблено на дезінформації, що тягне за собою визначення фейкових новин щодо достовірності їх змісту. Отже, стаття новин, яка містить просто неправильну інформацію (всупереч факту), розглядається як екземпляр фейкового класу (неправда), а стаття новин, що містить перевірену інформацію, є зразком справжніх новин (правда).

Стратегія збору даних для створення системи виявлення фейкових новин залежить від визначення, яке ви приймаєте для завдання. У більшості попередніх досліджень випадки фейкових новин збиралися зі списку підозрілих веб-сайтів. Відносно велика колекція такого типу — це набір даних із приблизно 20 000 новинних статей, зібраних у 2017 році [18]. Ці дані містять

тексти, отримані від восьми видавців новин, розділених на чотири класи: пропаганда (The Natural News і Activist Report), сатира (The Onion, The Borowitz Report і Clickhole), містифікація (American News and DC Gazette) і довірена (Gigaword News). Цей набір даних збалансований між класами та розділений на набори для навчання, перевірки та тестування. Однак шумна стратегія маркування всіх статей видавця на основі його репутації призведе до упередження класифікатора, навченого на цих даних, обмежуючи його здатність відрізняти окремі правдиві статті новин від випадків дезінформації. Іншими словами, дані, зібрані таким чином, не підходять для вивчення мовних моделей для визначення обману. Це радше допоможе розрізнити загальний стиль написання групи новинних веб-сайтів (відрізнити чутки від клікбейтів).

Для того, щоб побудувати систему класифікації текстів, щоб виявляти неправдивий вміст від правдивого на основі мовних ознак, нам потрібні статті новин, оцінені індивідуально та позначені відповідно до рівня правдивості. Цей тип збору даних є трудомістким, оскільки передбачає перевірку фактів для кожної новинної статті. Різноманітні веб-сайти для перевірки фактів здійснюють цей аналіз. Таким чином, один із способів зібрати дані про чутки та неправдиві новини — скористатися перевагами цих веб-сайтів і спробувати автоматично зібрати інформацію, таку як правдиві та неправдиві заголовки та їхні джерела.

### 1.3.1 MisInfoText: Репозиторій з маркованими текстами новин

Щоб усунути брак даних за допомогою надійних міток, було створено сховище текстів новинних статей, які були позначені веб-сайтами для перевірки фактів. Цей репозиторій містить три категорії даних:

- Посилання на всі загальнодоступні набори даних новин, які містять текст новинних статей та відповідні їм маркери правдивості. Це

полегшує як теоретичні, так і прикладні дослідження фейкових новин і автоматичне виявлення дезінформації.

- На додаток до наборів даних, опублікованих у попередніх дослідженнях, було використано скрейпінг поверх наборів даних, які містять твердження з позначкою правдивості та URL-адреси їхніх джерел, але не обов'язково текст статей новин. Наприклад, було знайдено два набори даних посилань із мітками правдивості в репозиторії Buzzfeed News. Ці посилання стають корисними для пошуку новинних статей, які вже були оцінені за фактичним змістом.
- Нарешті, даний репозиторій підтримує та використовує список потенційних веб-сайтів для перевірки фактів, щоб збирати більші обсяги даних. Збираючи дані безпосередньо з веб-сайтів перевірки фактів, таких як Snopes, репозиторій застосовує комбінацію автоматичних і ручних процедур. Наразі власники репозиторію зіскрібали весь архів веб-сайтів Snopes, Politifact та Emergent, а потім переходили за посиланнями, згаданими в кожній статті перевірки фактів на цих веб-сайтах, до джерел обговорюваних чуток. Ця категорія даних потребує ручної перевірки щоб переконатися, що текст дійсний і він фактично підтверджує обговорюване твердження.

### 1.3.2 Колекція даних Buzzfeed

Перше джерело інформації, яке використовується для збору повних статей новин із мітками правдивості в подібних дослідженнях, — це медіа-компанія Buzzfeed. Дана компанія опублікувала колекцію посилань на дописи у Facebook, спочатку зібрану для дослідження навколо виборів у США 2016 року [20]. Кожну URL-адресу в цьому наборі даних було надано експертам-людям, щоб вони могли оцінити кількість неправдивої інформації, що міститься у статті, на яку існують посилання. Посилання були зібрані з дев'яти сторінок

Facebook (три республіканських, три ліберальних і три з найпопулярніших видавців). Отриманий набір даних містить загалом 1380 статей новин на конкретну тему (вибори та кандидати в США).

Маркери правдивості мають чотиристоронню класифікаційну схему, яка включає 1090 переважно правдивих, 170 суміші правдивих і хибних, 64 переважно неправдивих і 56 статей, які не містять фактичного змісту. Інша цікава колекція URL-адрес, опублікованих Buzzfeed News, вказує на топ-50 фейкових новин у 2017 році. На відміну від набору даних про вибори в США, ці дані містять лише неправдиві новини зі статтями на різноманітні теми.

### 1.3.3 Колекція даних Snopes

Іншим популярним джерелом тренувальних даних для нейронних мереж є Snopes. Це добре відомий веб-сайт, який розвінчує чутки, керований командою експертних редакторів. Окрім пошуку чуток та згадок про розповсюдження веб-сайтів, Snopes надає докладні пояснення чуток та їх наслідків. Архів складається з багатьох сторінок перевірки фактів. На кожній сторінці Snopes обговорює твердження, цитує джерела (статті новин, форуми чи соціальні мережі, де було поширено твердження) і надає мітку правдивості твердження.

Вся скомпільована база даних Snopes містить приблизно 4000 рядків, кожен із яких містить твердження, обговорюване анотаторами Snopes, призначену йому позначку правдивості та текст статті новин, пов’язаної з претензією. Основна проблема використання цих даних для навчання/тестування детектора фейкових новин полягає в тому, що деякі з посилань на сторінці Snopes, які збираються автоматично, насправді не вказують на обговорювану статтю новин, тобто на джерело твердження. Багато посилань містять сторінки, які надають контекстну інформацію для перевірки фактів твердження. Тому не всі тексти в даному автоматично вилученому наборі даних є надійними або просто «допоміжним» джерелом твердження.

### 1.3.4 Тематичний поділ тренувальних даних

У 2019 році було проведено дослідження [10] вищевказаних репозиторіїв, щоб вивчити дані, які було зібрано з веб-сайтів перевірки фактів, і зрозуміти, які типи статей новин висвітлюються в них. Питання тем важливе, оскільки навчальні набори даних, які перекошені з точки зору теми, призведуть до класифікаторів, які не зможуть узагальнити рішення для відмінних від тренувальних тем. Загалом, дослідження на сьогоднішній день не вивчили, які теми з більшою чи меншою ймовірністю будуть представлені у фейкових новинах, хоча здається, що новини про політику, навколишнє середовище та здоров'я переважають. Дане дослідження показало, що що фейкові новини не встановлюють порядок денний для основних ЗМІ, вони тісно пов'язані з партизанськими новинами, беручи підказки з партійних сайтів щодо того, які типи тем і історій висвітлюються у фейкових новинах.

Щоб побудувати тематичну модель, попередньо було оброблено документи в навчальному наборі Рашкіна [18] (шляхом токенізації, нормалізації та видалення знаків пунктуації та стоп-слів) і подано езультуючі вектори документа в латентну модель розподілу Діріхле в бібліотеці Python Gensim. Було налаштовано кількість тем, щоб кожна тема представляла чітку категорію новин, не надто тонку чи грубу для візуального дослідження. Остаточна кількість тем, які дали найбільш чіткі результати, становила 10. На рисунку (рис. 1.5) показано хмари слів, які було отримано з 10 найважливіших слів у кожній темі, з їх вагою, представленим розміром шрифту.

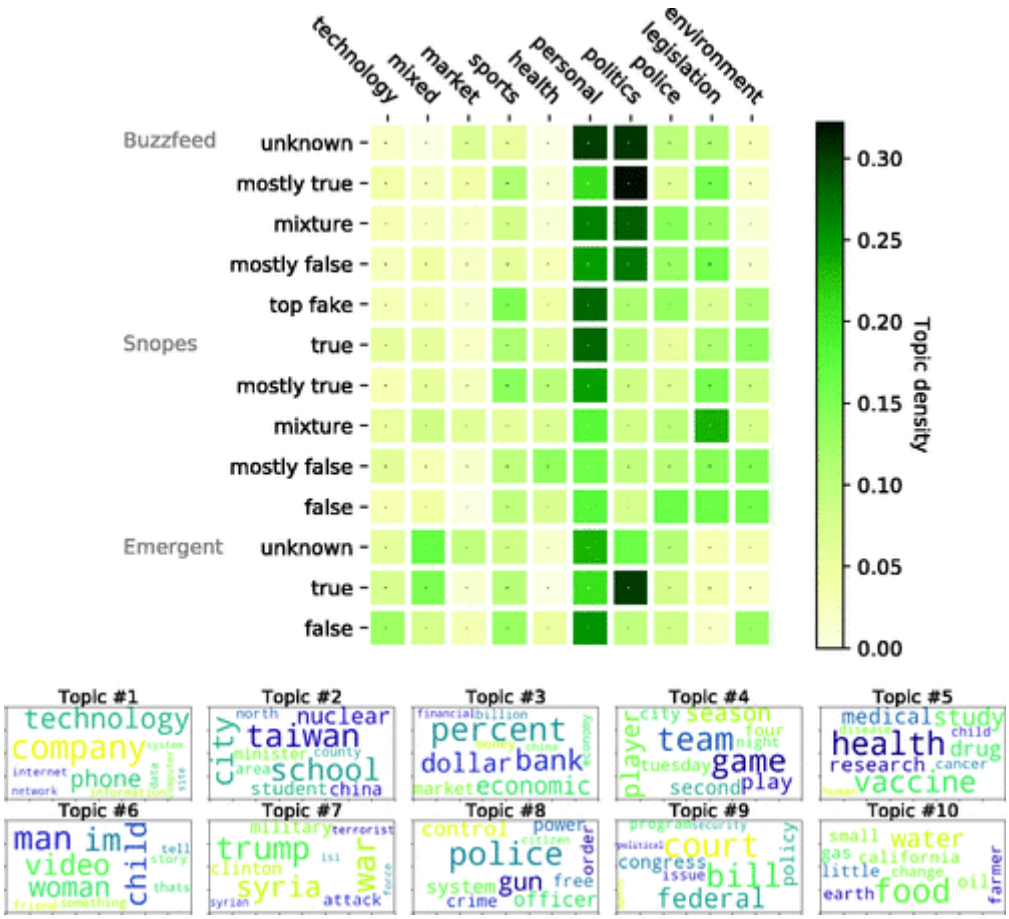


Рисунок 1.5 – Тематичний поділ тренувальних даних

Набір даних Buzzfeed (1380 статей), який здебільшого зосереджений на новинах, пов’язаних з виборами в США 2016 року, виявляється найменш різноманітним набором даних. Це було очікувано, оскільки цей набір даних охоплює теми виборів, особисті історії (кандидатів у президенти) та інші політичні теми, такі як історії, пов’язані з поліцією та системою законодавства. Набір даних Snopes (145 статей) є відносно різноманітнішим: на додаток до політичних тем, він містить деякі новини про спорт, навколишнє середовище та здоров’я. Зверніть увагу, що головна колекція фейкових новин Buzzfeed (33 статті) має більш подібний розподіл до колекції Snopes, і це тому, що Buzzfeed насправді зібрав цей набір даних, переглянувши веб-сайти Snopes і Politifact. Нарешті, набір даних Emergent (1612 статей) виділяється як найрізноманітніша колекція. Цей набір даних також відносно більший, що може опосередковано сприяти різноманітності тем. Хоча три набори даних разом охоплюють різноманітні новини, здається, що історії на певні теми, такі як ринок (економіка) та технології, представлені в цих колекціях менше.

Уважно розглянувши кожен рядок теплової карти (рис. 1.5), ми також виявимо, що деякі теми частіше зустрічаються в неправдивих новинах, ніж у правдивих. Наприклад, у наборі даних Snopes тема поліції частіше зустрічається у неправдивих статтях новин. У наборі даних Emergent теми технології та навколишнього середовища частіше зустрічаються у фальшивих новинах, тоді як для теми політики спостерігається протилежна картина. Ці відмінності можуть свідчити про притаманну різницю між дезінформацією та реальними новинами, або вони можуть просто означати, що досліджувані веб-сайти перевірки фактів упереджені до певних типів історій. Особисті історії, зокрема, часто з'являються у всіх наборах даних і на всіх маркерах правдивості. Ця тема є особливо цікавою, оскільки вона дійсно може бути послідовною ознакою новин типу чуток, але не обов'язково ознакою дезінформації.

# 2 МАТЕМАТИЧНА МОДЕЛЬ ГЕНЕРУВАННЯ ШЛЯХУ РОЗПОВСЮДЖЕННЯ ДЕЗІНФОРМАЦІЇ

Розглянувши загальні принципи поширення дезінформації, було визначено першочергову задачу наукової роботи: пошук першоджерела підробки новини та шляхи їх поширення. Щоб її вирішити, потрібно розв’язати наступні завдання:

- Знайти тренувальні дані
- Провести класифікацію
- Виконати первинну обробку текстів
- Застосувати алгоритм визначення подібності двох текстів
- Побудувати дерево поширення фейкових новин

Для виконання завдання було використано тренувальні дані з українського сайту StopFake.org. Вміст сайту складається із фейкові новини з різних ресурсів, теми новин і дати їх створення. Після успішного збереження тренувальних даних було проведено первинну обробку нових текстів.

## 2.1 Первинна обробка тренувальних даних

На основі дослідження [18] очевидною є потреба в первинній обробці тексту, для підвищення точності розрахунку схожості новин. Для отримання кінцевих тренувальних даних, обробка тексту поділяється на кілька етапів:

- Нормалізація.
- Приведення до нижнього регістру.
- Лематизація.
- Скорочення слів до основи.
- Видалення стоп-слів.
- Видалення шуму.

Щоб краще зрозуміти логіку вихідних даних потрібно розглянути детально кожен етап попередньої обробки.

Однією з найпростіших і найпоширеніших операцій з обробки тексту є приведення слів до нижнього регістру. Ця операція застосовується чи не у всіх дослідженнях, пов'язаних з обробкою природньої мови. Вона допомагає з не великими наборами даних і суттєво збільшує узгодженість очікуваного результату.

Видалення стоп-слів. На даному етапі всі слова, які не несуть ніякого змісту – видаляються. Ці слова включають:

- Спілки та союзні слова.
- Займенники.
- Прислівники.
- Частинки.
- Вигуки.
- Цифри і числівники.
- Розділові знаки і спеціальні символи (., - _ = + /!;:%? *).
- Вступні слова.
- Окремо стоять літери.
- Невизначені частки, прислівники і деякі звичайні прислівники.
- Слова-підсилювачі.
- Ряд деяких іменників, дієслів, прислівників.

Процес нормалізації дозволяє вилучити з тексту граматичну інформацію (відмінок, числа, види і часи дієслів, прикметники, рід тощо).

З граматичних міркувань у документах можуть використовуватися різні форми слова, наприклад, «організовувати», «організовує» та «організовуючи». Крім того, існують сімейства деривaційно споріднених слів зі схожими значеннями, наприклад, демократія, демократичний та демократизація. У багатьох ситуаціях здається, що було б корисно для пошуку за одним із цих слів повернути документи, які містять інше слово в наборі.

Метою стемінгу і лемматизації є приведення інфлективних форм слова, а інколи і деривaційно пов’язаних слів до спільної основи, незалежно від того наскільки одне слово відрізняється від іншого. Стимінг зазвичай відноситься до грубого евристичного процесу, який обриває кінці слів в досягти цієї мети, і часто включає видалення деривaційних афіксів.

Лемматизація зазвичай стосується традиційного мануального процесу із використанням словника та морфологічного аналізу слів, як правило, з метою видалення лише флективних закінчень і повернення основи або словникової форми слова, яка відома як лема. Якщо провести через даний процес слово «говорив» то стемінг скоріш за все поверне слово «говор» в той час як лемматизація поверне слово «говорити». Вони також можуть відрізнятися тим, що стемінг найчастіше згортає деривaційно пов’язані слова, тоді як лемматизація зазвичай згортає лише різні флективні форми леми. Лінгвістична обробка для стемінгу або лемматизації часто виконується за допомогою додаткового модуля до процесу індексування, і існує ряд таких компонентів, як комерційних, так і з відкритим кодом.

## 2.2 Методи класифікації дезінформації

Для класифікації дезінформації в текстових потоках тренувальних даних використовується сентиментальний аналіз попередньо обробленого тексту та подальша перевірка фактів за допомогою сервісів, описаних у розділі 1.2.7.

Першим кроком в алгоритмі виявлення дезінформації буде класифікація текстів за настроями. Мета цього процесу — визначити, до якого класу належить текст: негативного чи позитивного. Після отримання текстового класу можна відкинути частину текстів, які не мають негативних елементів, з чого можна зробити висновок про відсутність дезінформації.

Результатом первинної обробки тренувальних даних повинна стати вибірка ознак, що будується з гідно моделі Bag of words та метрики TF-IDF.

Модель мішка слів, або скорочено BoW (Bag of words), — це спосіб вилучення ознак з тексту для використання в моделюванні, за допомогою алгоритмів машинного навчання. Цей підхід дуже простий і гнучкий, його можна використовувати безліччю способів для вилучення ознак з документів. BoW – це представлення тексту, що описує присутність слів у документі. Воно включає дві речі:

- Словник відомих слів.
- Міра наявності відомих слів.

Його називають «мішком» слів, оскільки будь-яка інформація про порядок чи структуру слів у документі відкидається. Модель стосується лише того, чи зустрічаються відомі слова в документі, а не те, де в документі [21].

Ідея полягає в тому, що документи схожі, якщо вони мають схожий зміст. Крім того, лише із змісту ми можемо дещо дізнатися про значення документа. Пакет слів може бути простим або складним, в залежності від потреб дослідження. Складність полягає у вирішенні того, як створити словниковий запас відомих слів (або лексем), так і як оцінити наявність відомих слів.

Після того як словниковий запас вибрано, потрібно оцінити наявність слів у прикладах документів. Очевидним є дуже простий підхід: бінарне оцінювання наявності чи відсутності слів. Більше ефективні методи включають:

- Підрахунок кількості слів
- Визначення частоти появи кожного слова у тексті.

Проблема під час оцінки частоти слів полягає в тому, що дуже часті слова починають домінувати в документі (таким чином отримуючи більшу оцінку), але можуть не містити стільки ж «інформаційного вмісту» для моделі, як рідкісні, але, можливо специфічні для конкретної теми слова.

Один підхід полягає в тому, щоб змінити оцінку слів за частотою їх появи в усіх документах, щоб бали за часті слова які також часто зустрічаються в усіх документах, штрафувалися. Цей підхід до оцінки називається Term Frequency – Inverse Document Frequency, або TF-IDF де:

- Term Frequency (Частота терміну) – це оцінка частоти слова в поточному документі.

- Inverse Document Frequency (Інверсна частота в документі) – це оцінка того, наскільки рідко слово зустрічається в документах.

Таким чином досягається зважена оцінка словника, що не всі слова є однакова важливі чи цікаві. Така оцінка досягає ефекту висвітлення слів які є контекстуально чіткими та містять корисну інформацію в поточному документі.

Модель «мішок слів» дуже проста для розуміння та реалізації та пропонує велику гнучкість для налаштування конкретних текстових даних. Вона з великим успіхом використовується для задач прогнозування, таких як моделювання мови та класифікація документації. Тим не менш, він страждає від деяких недоліків, таких як:

- Словниковий запас: словниковий запас вимагає ретельного проектування, зокрема, щоб керувати розміром, що впливає на розрідженість представлення документів.
- Розрідженість: розріджені документи важче моделювати як з обчислювальних причин (складність пам’яті та часу), так і з інформаційних міркувань, коли проблема полягає в тому, щоб моделі використовували мало даних в великому просторі інформації.
- Значення: відкидання великої категорії слів ігнорує контекст і, у свою чергу, значення слів у документі (семантика). Контекст і значення можуть збагатити модель, можна було б розрізнити між тими самими словами, розташованими по-різному («це цікаво» проти «чи це цікаво»), синонімами («старий велосипед» проти «вживаний велосипед») і набагато більше.

Після отримання вибірки ознак проводиться класифікація тексту за настроями. Найпоширенішим методом такого парсингу тексту є градієнтний бустинг. Цей метод дуже часто зустрічається у дослідженнях зв’язаних з машинним навчанням де згадуються задачі регресії та класифікації. Він дозволяє створити сильну модель на основі багатьох слабких моделей. Ця технологія використовує дерево прийняття рішень.

Після визначення класів текстів відбираються тексти з негативним тоном і запускається процес перевірки фактів за допомогою сервісів фактчекінгу. Загальний алгоритм класифікації зображено на рисунку 2.1

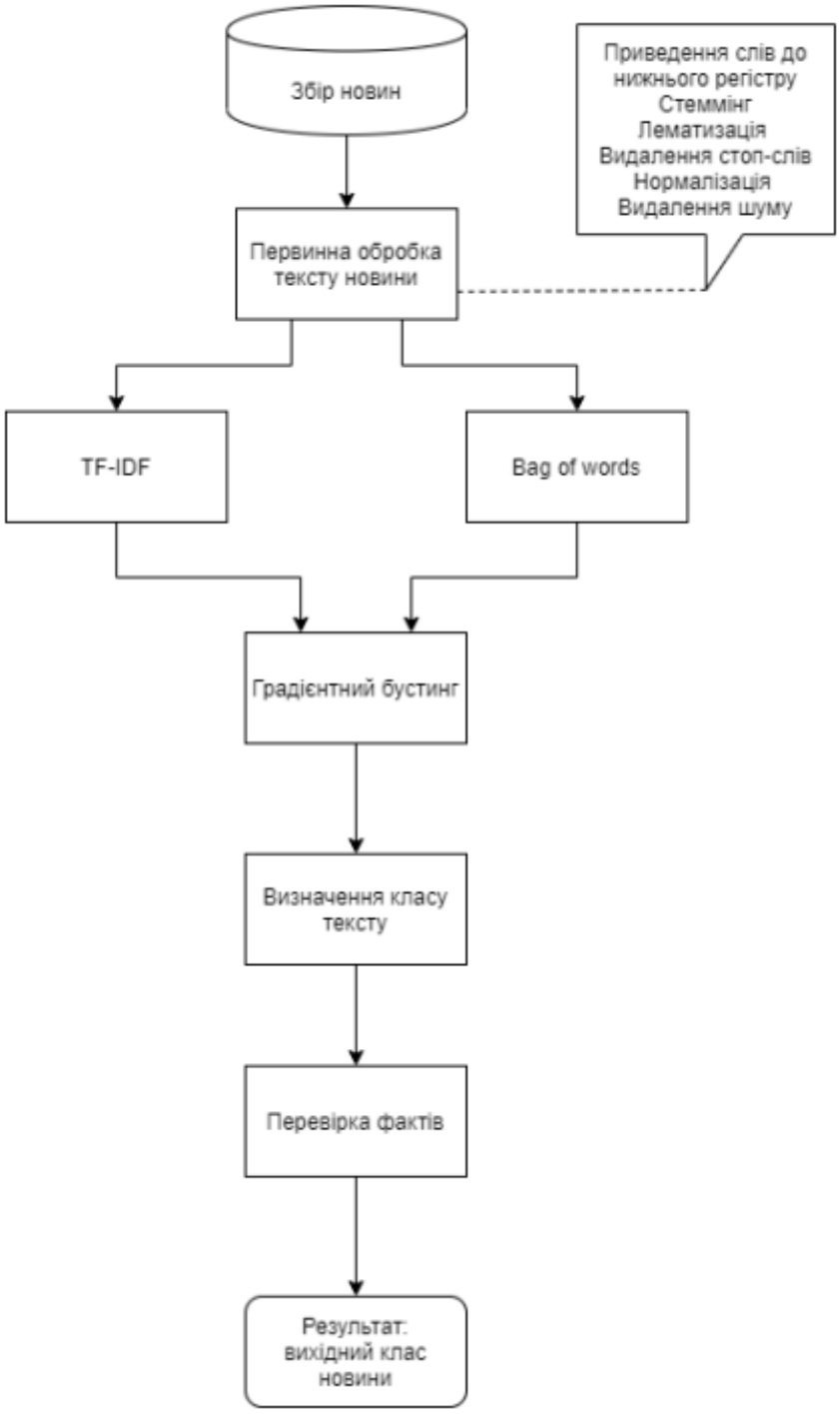


Рисунок 2.1 – Алгоритм класифікації тренувальних даних відповідно до наявності чи відсутності елементів дезінформації

## 2.3 Методи визначення подібності текстових документів

Після визначення новини яка належить до категорії дезінформації, її можна порівняти з іншими новинами і при достатньому відсотку подібності відповідно розмітити їх. Існує кілька способів визначення подібності текстів, у даній роботі було використано 2, а саме коефіцієнт Жаккара та косинус подібності. Для використання цих методів, потрібно провести попередню обробку тексту яку було описано в розділі 2.2.

Коефіцієнт Жакарда — це звичайне вимірювання близькості, яке використовується для обчислення подібності між двома об'єктами, такими як два текстові документи. Коефіцієнт Жаккара можна використовувати, щоб знайти подібність між двома асиметричними бінарними векторами або знайти подібність між двома множинами. У літературі подібність Жаккара, також може називатися індексом, коефіцієнтом, несхожістю і відстанню.

Даний коефіцієнт розраховується шляхом ділення кількості спостережень в обох множинах на кількість спостережень в одній з множин. Іншими словами, подібність Жаккара можна обчислити як розмір перетину, поділений на розмір об'єднання двох множин. Це можна записати в нотації множин, використовуючи перетин і об'єднання двох множин:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

В даному вигляді перетин вказує на кількість елементів спільних для обох множин, а об'єднання кількість елементів в обох множинах (спільних і різних). Коефіцієнт буде рівним 0 якщо обидві множини не мають спільних значень і 1 якщо множини ідентичні.

Для кращого розуміння даної метрики варто привести практичний приклад. Для прикладу було взято 2 фейкові новини на спільну тему (рис. 2.2 та 2.3)

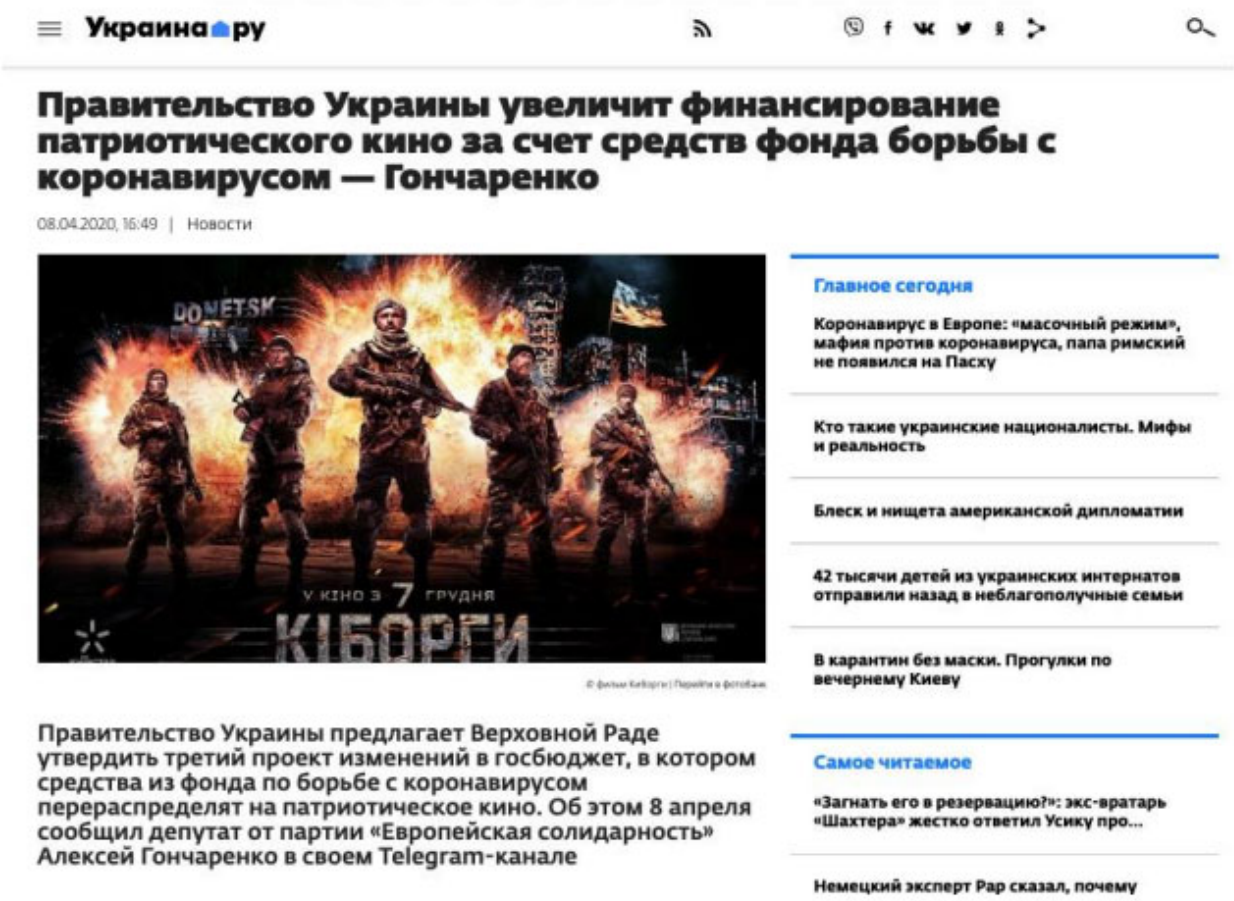


Рисунок 2.2 – Фіктивна новина на веб-сайті Ukraine.ru


Рисунок 2.3 – Фіктивна новина на веб-сайті Strana.ua

Для обчислення коефіцієнта Жаккара потрібно провести попередню обробку тексту для отримання двох множин елементів з потенційною дезінформацією. Результуючі елементи показано в діаграмі Венна (рис. 2.4)

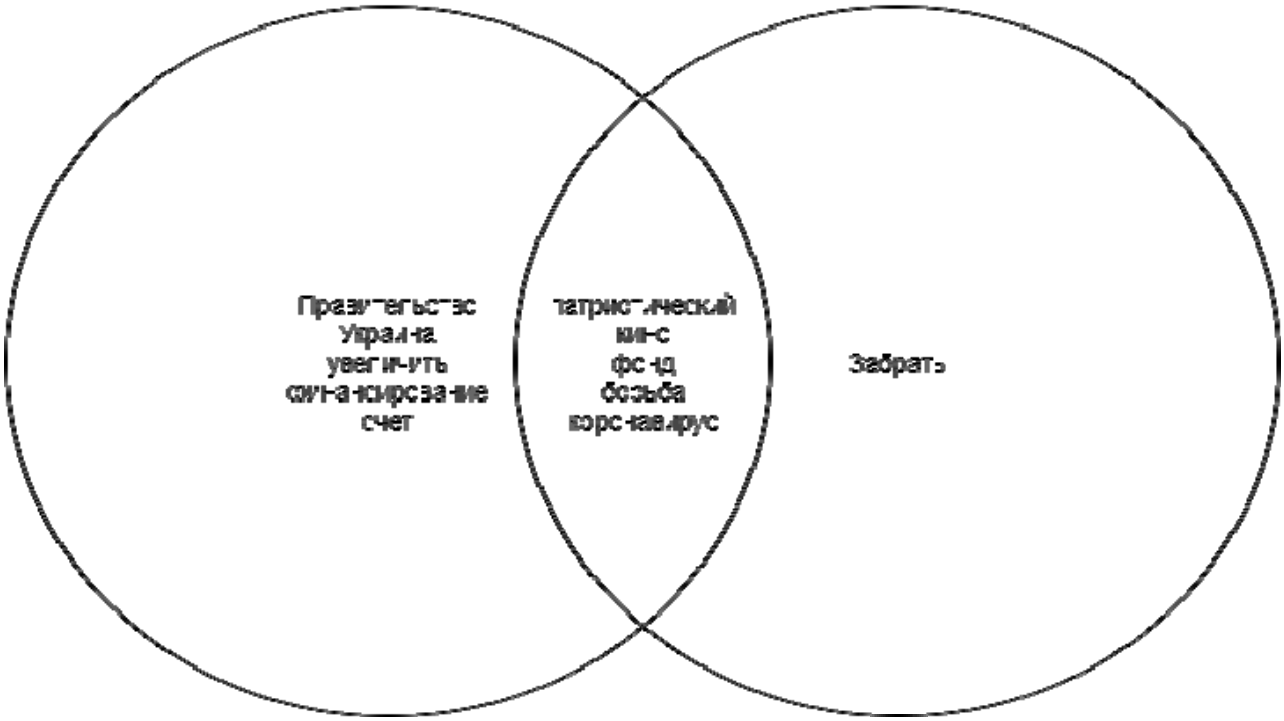


Рисунок 2.4 – Діаграма Венна з елементами обох фіктивних новин

Маючи попередньо оброблені дані можна обчислити коефіцієнт Жаккара:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{5}{5+5+1} = 0.45$$

Отриманий результат дає змого з великою долею впевненості вважати що обидві новини було написано на спільну тематику.

Інша метрика, яку було використано в даній роботі, відома як косинус подібності. Косинус подібності — це показник, який використовується для визначення схожості документів незалежно від їх розміру.

Математично він вимірює косинус кута між двома векторами, спроектованими в багатовимірному просторі. У цьому контексті два вектори, про які йде мова, є масивами, що містять кількість слів двох документів. При нанесенні на багатовимірний простір, де кожен вимір відповідає слову в

документі, косинусна подібність фіксує орієнтацію (кут) документів, а не величину.

Косинус подібності є корисною метрикою, тому що навіть якщо два подібних документи знаходяться далеко один від одного на евклідову відстань через розмір (наприклад, слово «цвіркун» з'явилося 50 разів в одному документі та 10 разів в іншому), вони все одно можуть мати менший кут між ними. Чим менший кут, тим вище подібність.

При використанні TF-IDF та BoW формула для обчислення косинусу подібності виглядає наступним чином:

$$\cos(\theta) = \frac{TFIDF_{news1} \cdot TFIDF_{news2}}{|TFIDF_{news1}||TFIDF_{news2}|}$$

Проведемо обчислення косинусу подібності для попередньо згаданих новин (рис. 2.2, 2.3). Першим кроком є побудова BoW для обидвох новин. Для зручності результат цього процесу приведений в таблиці 2.1

Таблиця 2.1 – BoW для обох новин

| | A | B |
|---|---|---|
| Привительство | 1 | 0 |
| Украина | 1 | 0 |
| увиличить | 1 | 0 |
| финансирование | 1 | 0 |
| патриотический | 1 | 1 |
| кино | 1 | 1 |
| счет | 1 | 0 |
| фонд | 1 | 1 |
| борьба | 1 | 1 |
| коронавирус | 1 | 1 |
| забрать | 0 | 1 |

Наступним кроком є нормалізація елементів використовуючи TF-IDF

Таблиця 2.2 – Нормалізація з використанням TF-IDF

| | A | B | TF-IDF 1 | TF-IDF 2 |
|---|---|---|---|---|
| Привительство | 1 | 0 | 0,09091 | 0 |
| Украина | 1 | 0 | 0,09091 | 0 |
| увиличить | 1 | 0 | 0,09091 | 0 |
| финансирование | 1 | 0 | 0,09091 | 0 |
| патриотический | 1 | 1 | 0,09091 | 0,09091 |
| кино | 1 | 1 | 0,09091 | 0,09091 |
| счет | 1 | 0 | 0,09091 | 0 |
| фонд | 1 | 1 | 0,09091 | 0,09091 |
| борьба | 1 | 1 | 0,09091 | 0,09091 |
| коронавирус | 1 | 1 | 0,09091 | 0,09091 |
| забрать | 0 | 1 | 0 | 0,09091 |

Таблиця 2.3 – обчислення IDF обох новин

| | A | B | TF-IDF 1 | TF-IDF 2 | IDF 1 | IDF 2 |
|---|---|---|---|---|---|---|
| Привительство | 1 | 0 | 0,09091 | 0 | 1,6545 | 1,6525 |
| Украина | 1 | 0 | 0,09091 | 0 | 1,6545 | 1,6525 |
| увиличить | 1 | 0 | 0,09091 | 0 | 1,6545 | 1,6525 |
| финансирование | 1 | 0 | 0,09091 | 0 | 1,6545 | 1,6525 |
| патриотический | 1 | 1 | 0,09091 | 0,09091 | 1 | 1 |
| кино | 1 | 1 | 0,09091 | 0,09091 | 1 | 1 |
| счет | 1 | 0 | 0,09091 | 0 | 1,6545 | 1,6525 |
| фонд | 1 | 1 | 0,09091 | 0,09091 | 1 | 1 |
| борьба | 1 | 1 | 0,09091 | 0,09091 | 1 | 1 |
| коронавирус | 1 | 1 | 0,09091 | 0,09091 | 1 | 1 |
| забрать | 0 | 1 | 0 | 0,09091 | 1,6545 | 1,6525 |

Таблиця 2.4 – Результуючий IDF для обох новин

| IDF 2 | TF-IDF 1 |
|---|---|
| 1,6525 | 0,15093 |
| 1,6525 | 0,15093 |
| 1,6525 | 0,15093 |
| 1,6525 | 0,15093 |
| 1 | 0,09091 |
| 1 | 0,09091 |
| 1,6525 | 0,15093 |
| 1 | 0,09091 |
| 1 | 0,09091 |
| 1 | 0,09091 |
| 1,6525 | 0 |

Маючи результати попередньої обробки даних (табл. 2.4) можна обчислити косинус подібності:

$$\cos(\theta) = \frac{TFIDF_{news1} \cdot TFIDF_{news2}}{|TFIDF_{news1}||TFIDF_{news2}|} = \frac{0,041}{0,39 \cdot 0,25} = 0,41$$

Схожість двох новин за косинусом подібності рівна 41%.

Подібність Жаккарда приймає лише унікальний набір слів для кожного речення/документа, тоді як косинусна подібність приймає загальну довжину векторів. (ці вектори добре працюють з використанням BoW і tf-idf). Це означає, що при повторенні слова «друг» у реченні кілька разів, косинусна подібність зміниться, а подібність Жаккара — ні. Подібність Жаккарда хороша для випадків, коли дублювання не має значення, косинусна подібність хороша для випадків, коли дублювання має значення під час аналізу подібності тексту.

Виходячи з вищевказаних висновків в даній роботі для визначення схожості новин було використано косинус подібності.

## 2.4 Визначення шляху розповсюдження дезінформації

Однією з основних задач наукової робити є дослідження та визначення шляху розповсюдження дезінформації. Використовуючи метод визначення схожості новин розглянутий в розділі 2.3, було створено алгоритм пошуку подібних фіктивних новин використовуючи косинус подібності.

Для візуалізації розв’язку даної задачі було створено псевдокод та приведено алгоритм (рис. 2.5) результатом якого є граф розповсюдження дезінформації з ієрархією відповідно до часових міток.

```
queryNews = 'text'
queryNewsVector = tfidf(queryNews.text)
similarNews = []
for news in newsArray:
 currentNewsVector = tfidf(news.text)
 if similarity(queryNewsVector, currentNewsVector) > 0.6:
 similarNews.append(news)
similarNews.sort(key=lambda news: news.date, reverse=True)
generateGraph()
```

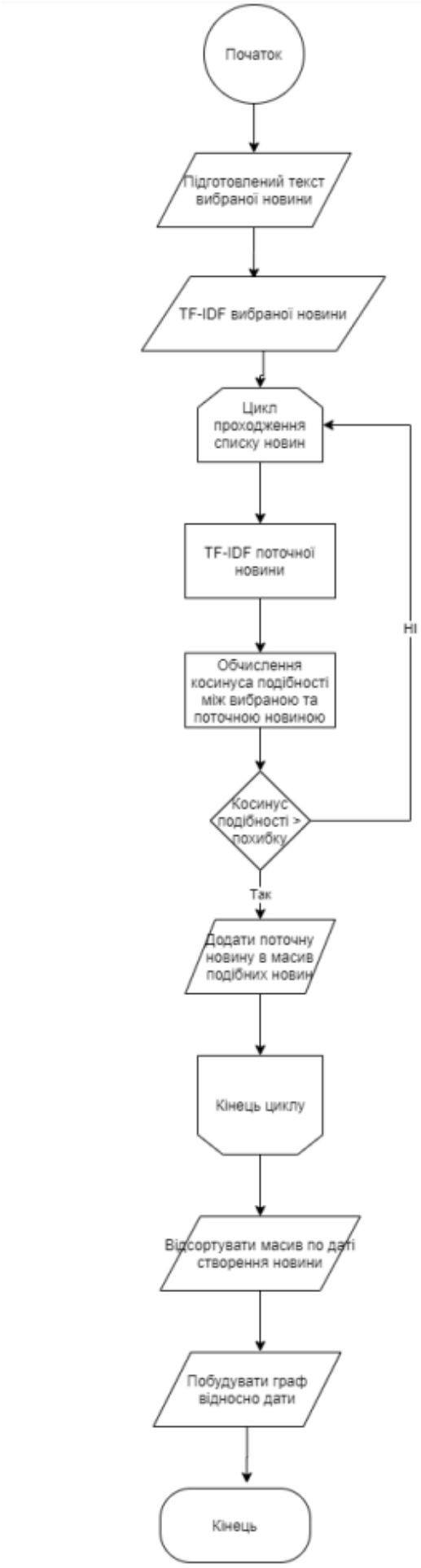


Рисунок 2.5 – Алгоритм створення графу шляху дезінформації

Для наочного прикладу роботи алгоритму було використано фейкову новину взяту з архіву дезінформації (рис. 2.6)


Рисунок 2.6 – Фіктивна новина оригінально опублікована на веб-сайті Agronews

Використовуючи косинус подібності, було знайдено схожі статті (рис. 2.7, 2.8) та на їх основі сформовано таблицю 2.5.

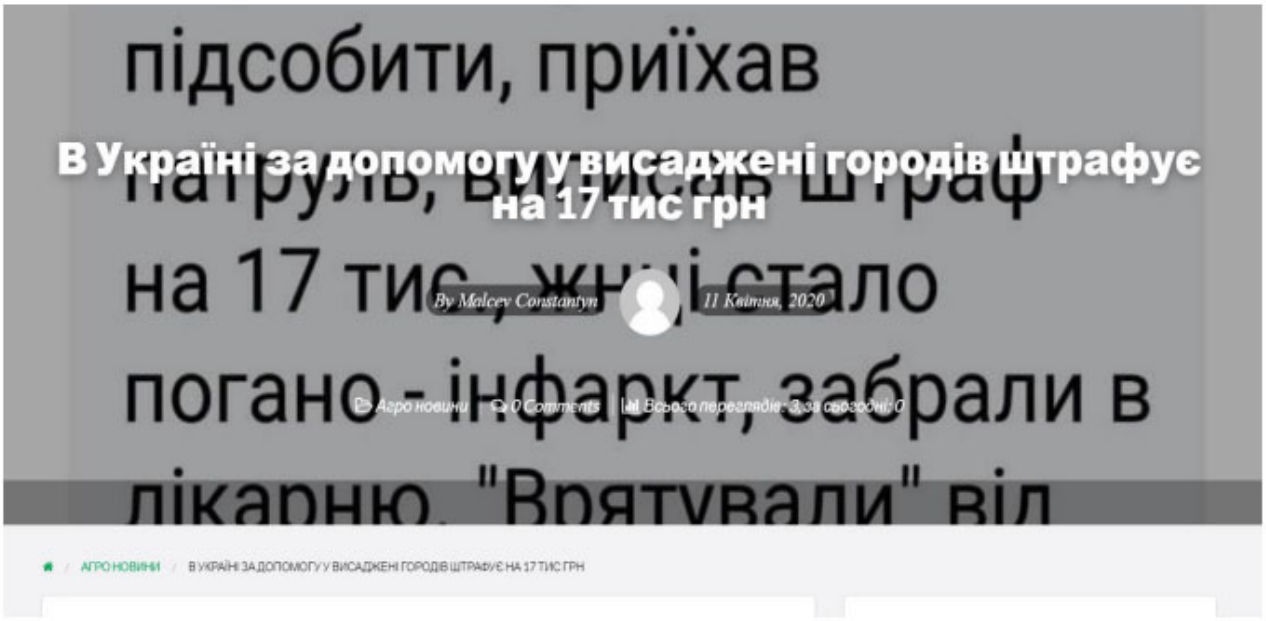


Рисунок 2.7 – Фіктивна новина оригінально опублікована на веб-сайті Unicorns


Рисунок 2.8 – Фіктивна новина оригінально опублікована на веб-сайті Volyninfo

Таблиця 2.5

| Веб-сайт | Оригінальний час публікації | Косинус подібності |
|---|---|---|
| Agronews | 11.04.2020, 16:14 | - |
| Unicorns | 11.04.2020, 20:05 | 1 |
| Volyninfa | 12.04.2020, 12:04 | 0,8 |

На основі даних в таблиці 2.5 було згенеровано граф розповсюдження дезінформації (рис. 2.9)

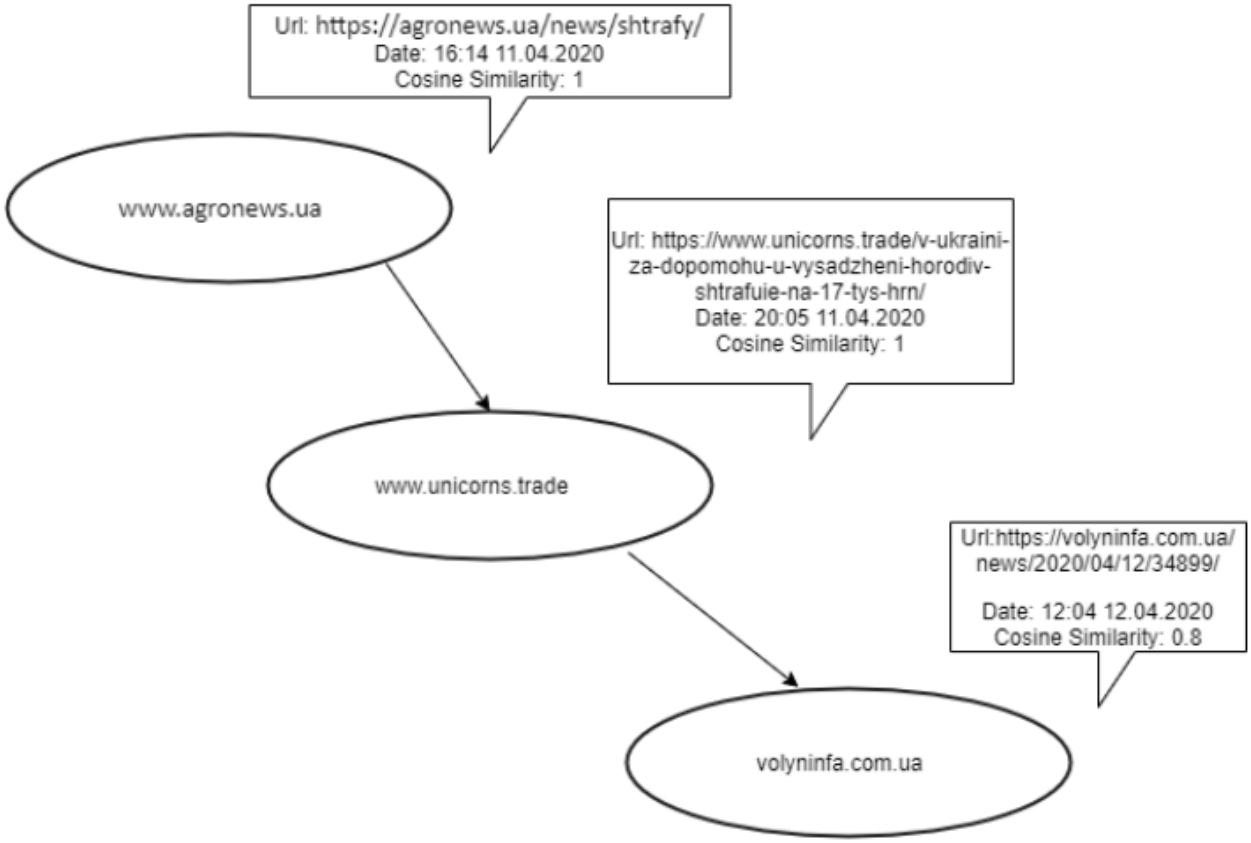


Рисунок 2.9 – Граф розповсюдження дезінформації

Кінцевий алгоритм розв’язання задачі визначення шляху розповсюдження дезінформації приведено на рисунку 2.10

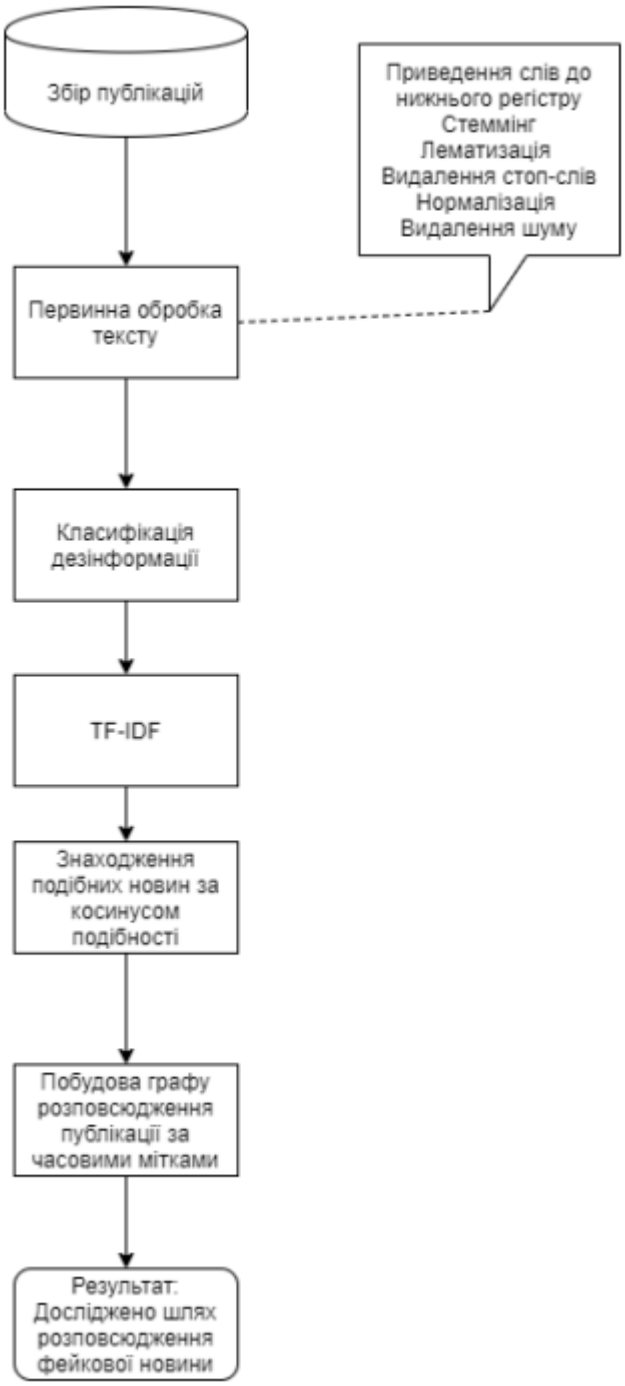


Рисунок 2.10 – Кінцевий алгоритм

# 3 ПРОГРАМНА РЕАЛІЗАЦІЯ

## 3.1 Вимоги до програмного забезпечення

Для виконання поставленого завдання даної роботи та відповідно до аналізу актуального ПЗ, для реалізації програмного забезпечення для виявлення та генерації шляху поширення дезінформації потрібно:

- Використовуючи фреймворки парсингу української та російської мов, розробити класифікатор виявлення дезінформації, а також генератор шляху поширення фіктивних новин.
- Програмно реалізувати вищеописані алгоритми: первинна обробка тексту, аналіз настрою тексту, створення BoW з використанням TF-IDF, косинуси подібності, побудова графу.
- Реалізувати інтерфейс користувача ПЗ
- Протестувати результуюче ПЗ

## 3.2 Архітектура ПЗ

Використовуючи попередньо описані дослідження, було змодельовано програмну архітектуру. Для її опису було використано такі UML-діаграми:

- Діаграма класів.
- Діаграма послідовностей.
- Діаграма розгортання.

Перед моделюванням архітектури важливо прийняти основні програмні рішення щодо використання фреймворків.

Однією з основних задач була підтримка української та російської мов, тобто для обробки тренувальних даних цих мов потрібен відповідний морфологічний аналізатор. Для української мови було використано pymorphy2

[22]. Даний фреймворк написаний мовою Python (працює під 2.7 та 3.5+). Він уміє:

- Приводити слово до нормальної форми (наприклад, "люди -> людина", або "гуляв -> гуляти").
- Ставити слово у потрібну форму. Наприклад, ставити слово у множину, змінювати відмінок слова тощо.
- Повертати граматичну інформацію про слово (число, рід, відмінок, частина мови тощо)

Під час роботи використовується словник OpenCorpora; для незнайомих слів будуються гіпотези. Бібліотека досить швидка: зараз швидкість роботи - від кількох тис слів/сек до > 100тис слів/сек (залежно від виконуваної операції, інтерпретатора та встановлених пакетів).

Для російської мови було використано бібліотеку Dostoevsky [23]. Для веб-інтерфейсу було використано фреймворк Angular [24]. Angular — це платформа розробки, побудована на TypeScript. Як платформа, Angular включає:

- Компонентний фреймворк для створення масштабованих веб-додатків.
- Колекцію добре інтегрованих бібліотек, які охоплюють широкий спектр функцій, включаючи маршрутизацію, керування формами, зв’язок клієнт-сервер тощо.
- Набір інструментів для розробників, які допоможуть розробляти, створювати, тестувати та оновлювати код.

Діаграма класів приведена на рисунку 3.1. Нижче розписано функціонал кожного класу та відповідних методів:

- NewsPreprocessor: відповідає за первинну обробку тексту. Використовує MorphAnalyzer бібліотеки pymorph2.
    - toUpperCase: приведення слів до нижнього регістру;
    - lemmatize: лематизація;
    - removeStopWords: видалення стоп-слів;
    - normalize: нормалізація;
    - noizeRemoval: видалення шуму;

- DesinformationClassifier: класифікація тексту за сентиментом. Використовує модель FastTextSocialNetworkModel бібліотеки Dostoevsky.
    - Classify: виконує класифікацію конкретного тексту за сентиментом.
- VectorsCalculator: реалізація алгоритмів для обчислення TF-IDF та косинуса подібності двох новин.
    - getTfidf: перетворює текст новини в векторний вигляд;
    - getCosine: обчислює косинус подібності двох новин;
- PathGenerator: основний клас для генерації шляху розповсюдження дезінформації. Реалізує основний алгоритм програмного застосунку. Метод generate повертає відсортований масив новин, який показує шлях розповсюдження дезінформації.

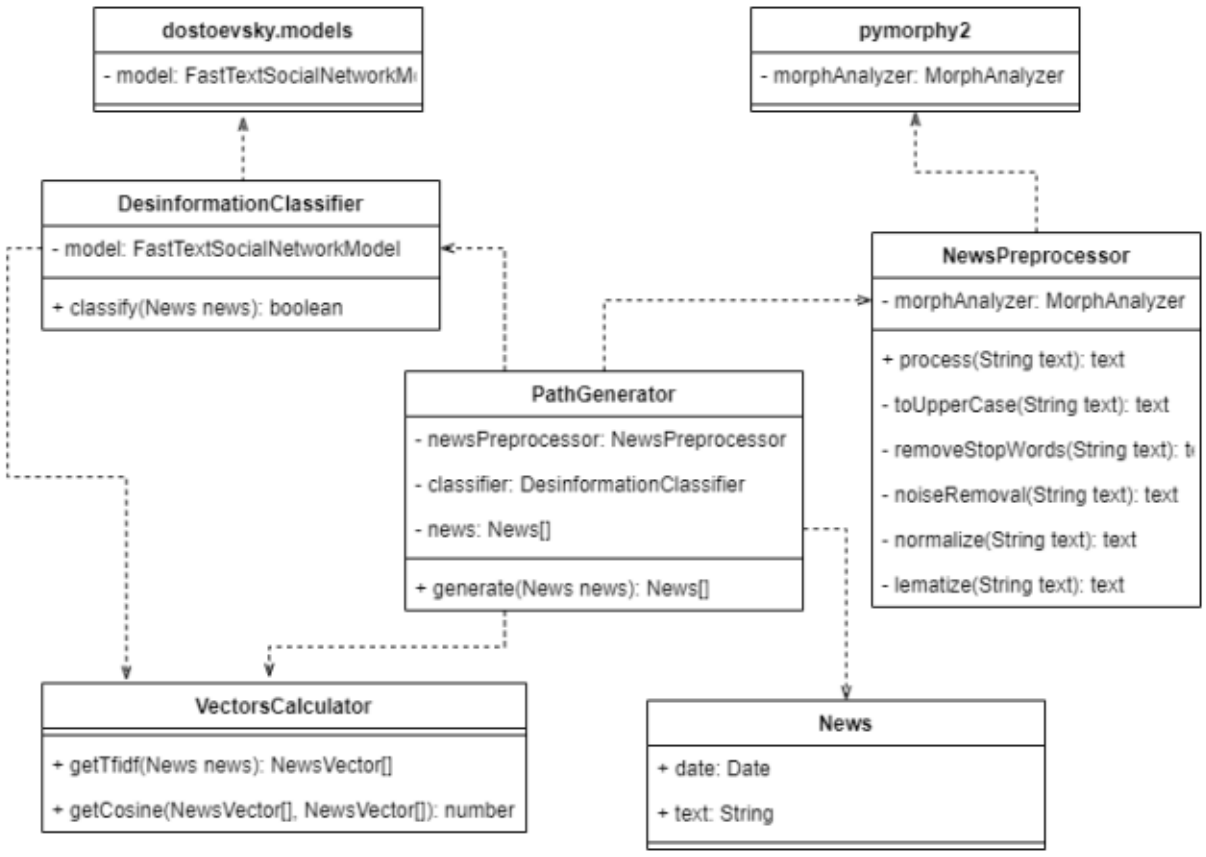


Рисунок 3.1 – UML-діаграма класів

На рисунках 3.2 та 3.3 наведено діаграми послідовностей для програмного вирішення основних задач проектованого ПЗ, а саме:

- Класифікація новин відповідно до наявності елементів дезінформації

- Генерація графу розповсюдження дезінформації

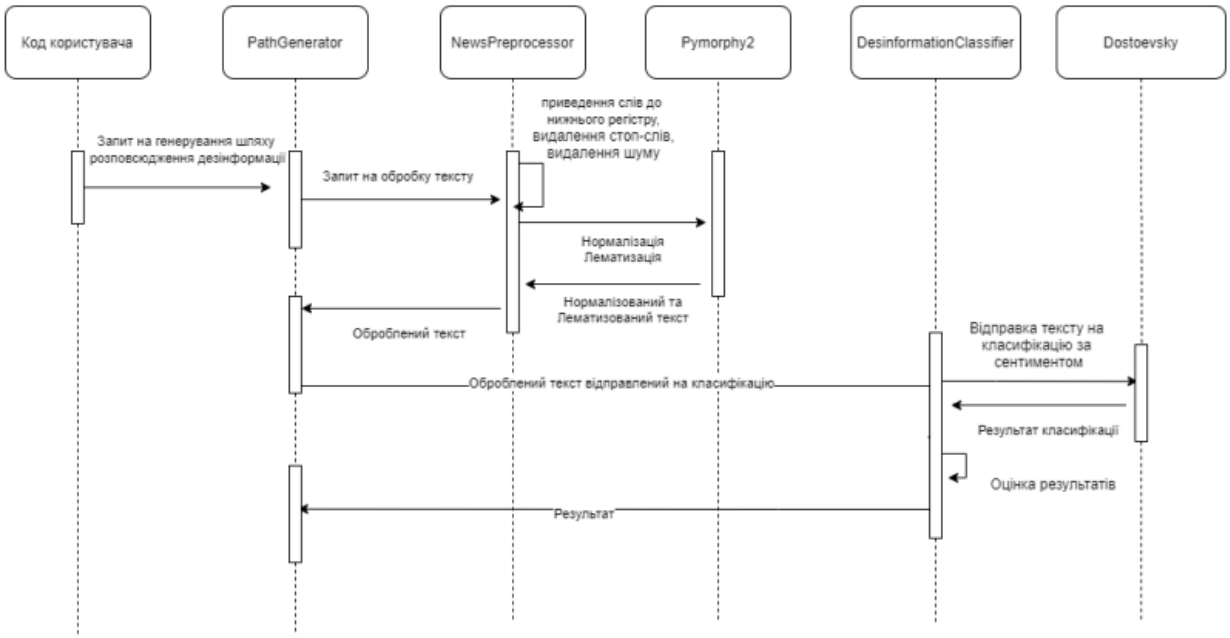


Рисунок 3.2 – UML-діаграма послідовностей для класифікації

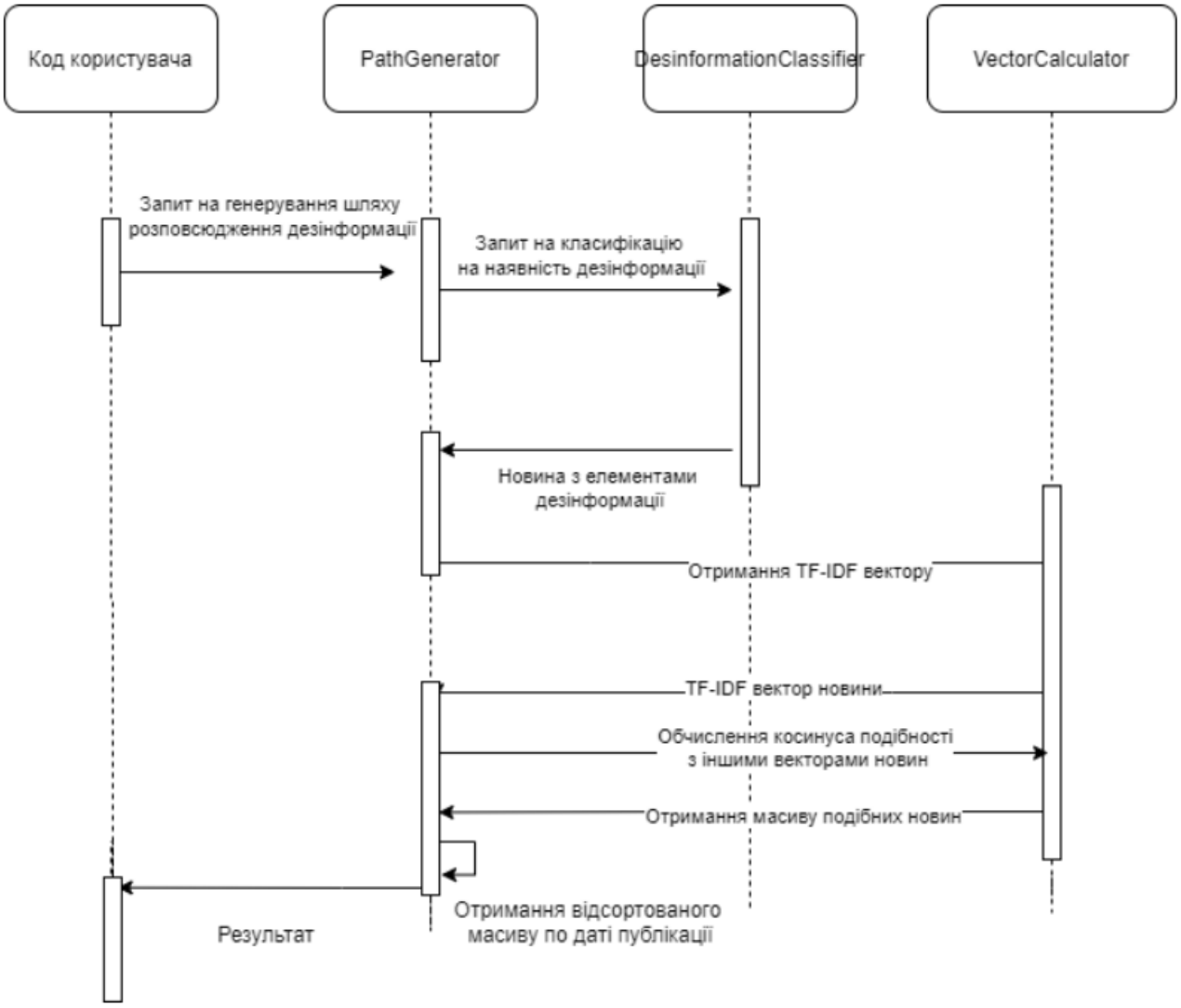


Рисунок 3.3 – UML-діаграма послідовностей генерації графу розповсюдження

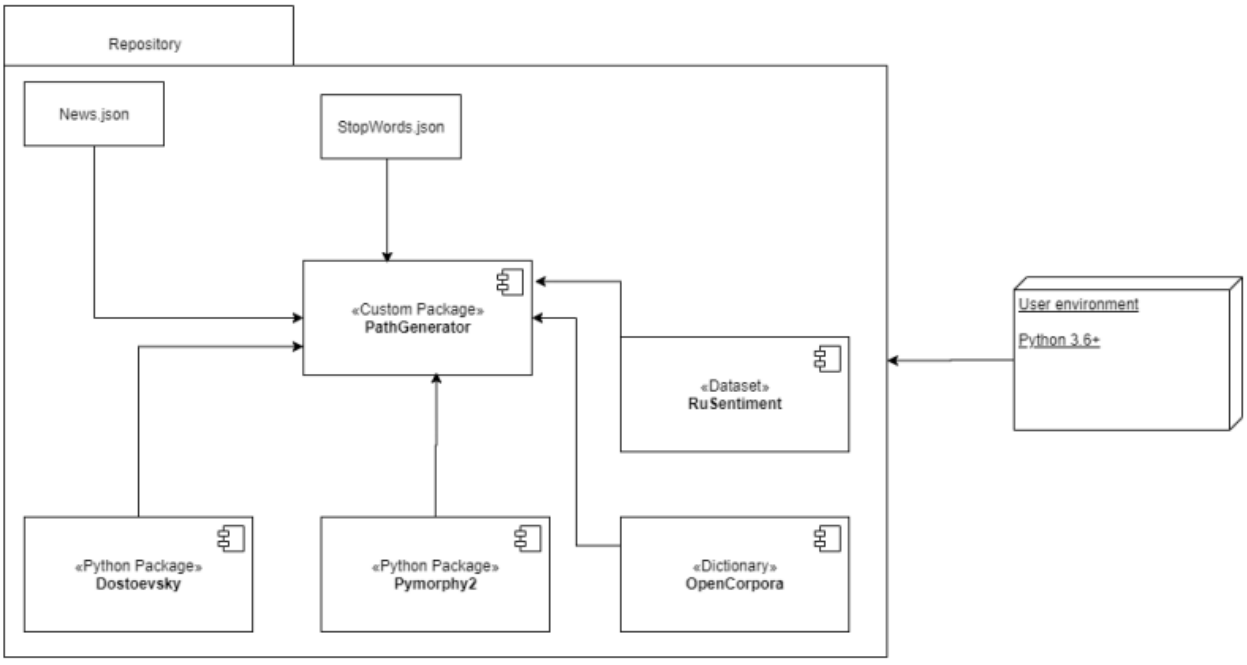


Рисунок 3.4 – UML-діаграма розгортання

## 3.3 Встановлення застосунку

Для встановлення та користування ПЗ користувач повинен мати інтерпретатор Python 3.6+, а також встановити всі пакети вказані на діаграмі розгортання (рисунок 3.4).

Для коректної роботи ПЗ у систему Windows 10 повинні бути присутні інструменти Visual Studio Build Tools. Запуск роботи серверу потребує консольної команди server.py.

Щоб встановити веб-інтерфейс потрібно встановити NodeJS використовуючи менеджер пакетів npm.

## 3.4 Використання застосунку

Застосунок було розроблено з орієнтацією на базу веб-користувачів, але його можна запустити і локально. Для цього потрібно запустити локальний сервер (рис. 3.5)

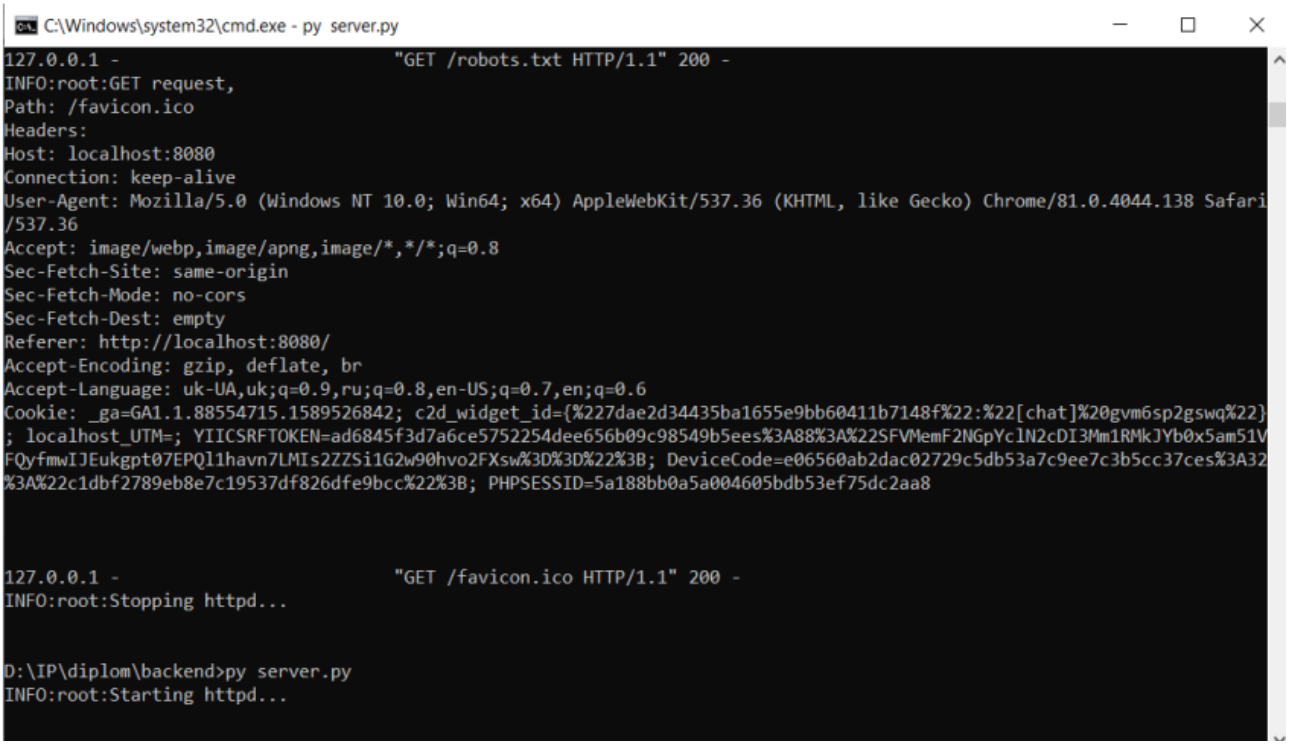


Рисунок 3.5 – Локальний запуск додатку

При коректній роботі, доступ до веб-інтерфейсу можна отримати перейшовши на домашню адресу локальної машини.

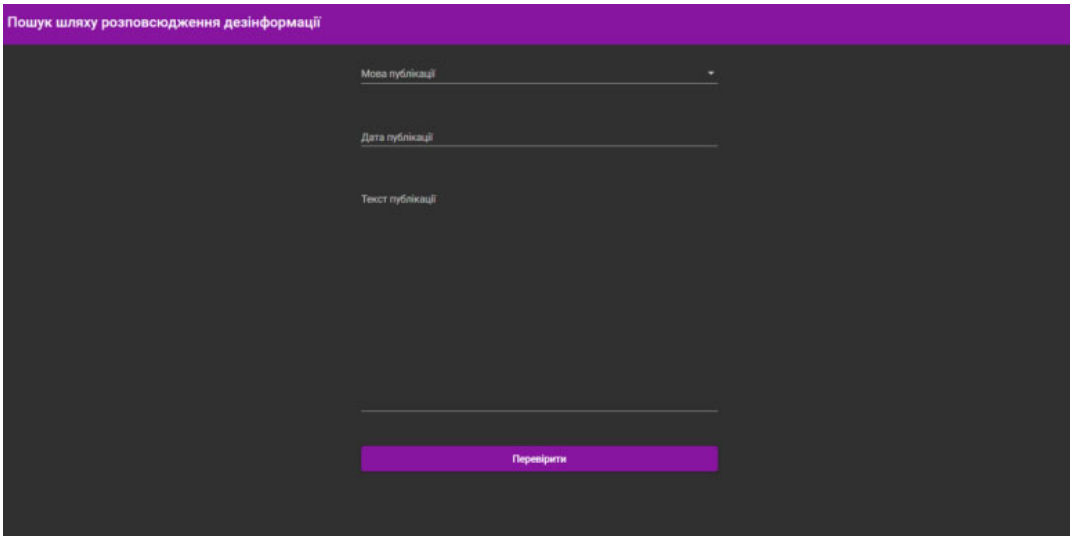


Рисунок 3.6 – Веб-інтерфейс додатку

Для перевірки правдивості тексту, користувачу необхідно обрати мову введеного тексту, ввести дату публікації, а також текст. Після цього натиснути кнопку перевірити і зачекати на результат (рис. 3.7).

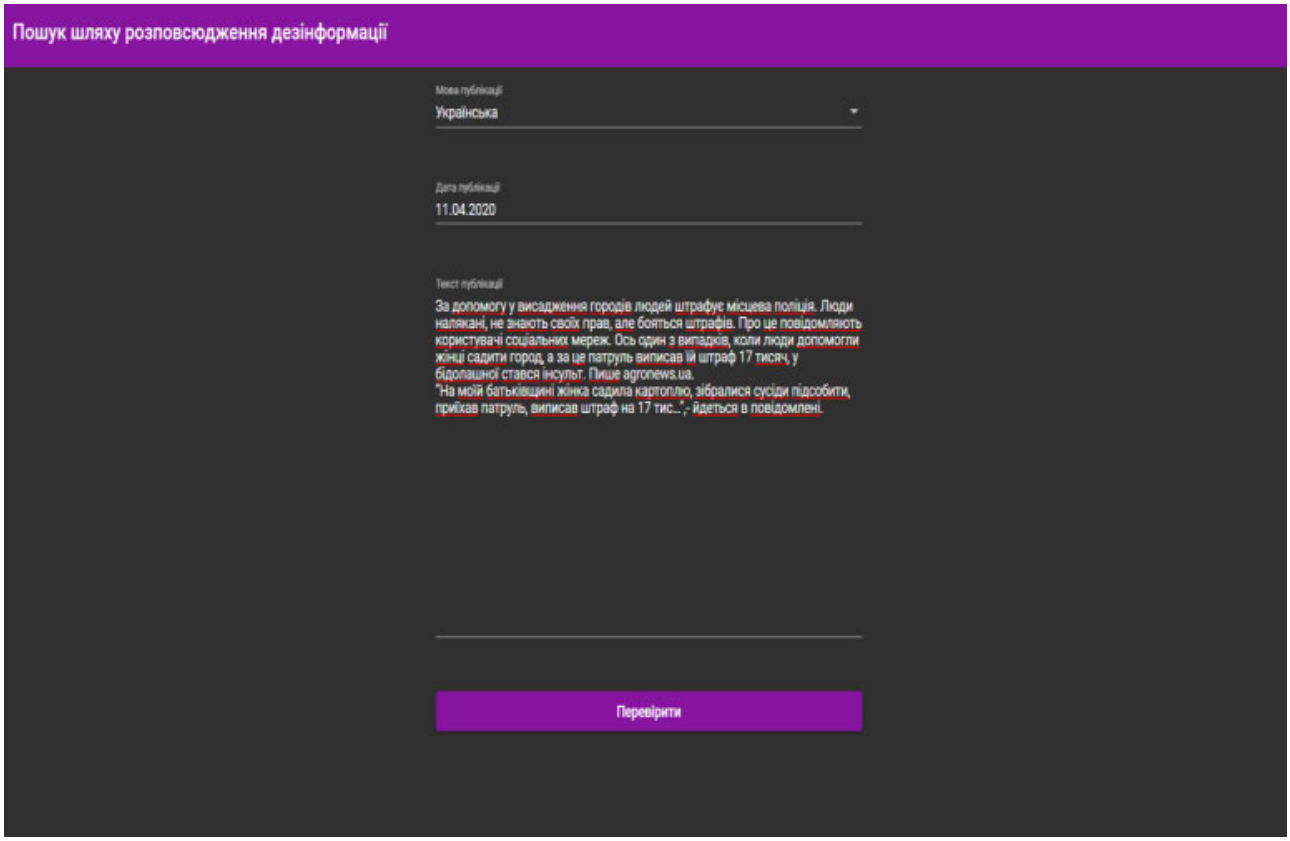


Рисунок 3.7 – Тестові вхідні дані

В результаті генерується граф показаний на рисунку 2.9

# 4 ОЦІНКА ЯКОСТІ РОБОТИ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Протягом розробки програмного забезпечення було використано і змінено кілька алгоритмів для класифікатора, тому у даному розділі було показано оцінку якості кожного з використаних методів.

## 4.1 Статистика класифікаторів та використаних метрик

Оскільки ми мали справу з незбалансованими даними, точність прогнозування була оманливим показником, оскільки вона відображає основні розподіли класів, що ускладнює роботу класифікатора щодо класу меншості. З цієї причини було використано оцінку F1 [25] та метрики запам’ятовування, оскільки чим вище значення, яке приймають ці показники, тим кращий клас, який нас цікавить.

У таблиці 4.1 показані показники оцінки для всіх розглянутих нами класифікаторів. Можна помітити, що показники класифікації залежать від типу класифікатора та від вилучених ознак, які використовуються для класифікації. Логістична регресія з моделлю BoW була найефективнішим класифікатором, коли ми зробили надмірну вибірку даних, досягнувши найвищого бала F1 (71%), за нею йшли наївна модель Байєса з моделлю BoW (70%) і SVM з TF-IDF (69%).

Таблиця 4.1 – Оцінка використаних метрик

| Re-sampling | Classifier | Pre-processing | Precision | Accuracy | Recall | $F_1$ score |
|---|---|---|---|---|---|---|
| Over | **Naïve Bayes** | **BoW** | **0.64** | **0.92** | **0.78** | **0.70** |
| | **Logistic regression** | **BoW** | **0.73** | **0.93** | **0.68** | **0.71** |
| | SVM | BoW | 0.73 | 0.93 | 0.63 | 0.68 |
| | SGD | BoW | 0.70 | 0.92 | 0.62 | 0.66 |
| | Random forest | BoW | 0.82 | 0.92 | 0.42 | 0.56 |
| | Naïve Bayes | TF-IDF | 0.38 | 0.82 | 0.88 | 0.53 |
| | Logistic regression | TF-IDF | 0.68 | 0.92 | 0.67 | 0.67 |
| | **SVM** | **TF-IDF** | **0.79** | **0.94** | **0.60** | **0.69** |
| | SGD | TF-IDF | 0.60 | 0.91 | 0.64 | 0.62 |
| | Random forest | TF-IDF | 0.76 | 0.93 | 0.58 | 0.66 |
| Under | Naïve Bayes | BoW | 0.23 | 0.64 | 0.91 | 0.37 |
| | Logistic regression | BoW | 0.38 | 0.83 | 0.79 | 0.51 |
| | SVM | BoW | 0.34 | 0.80 | 0.79 | 0.47 |
| | SGD | BoW | 0.26 | 0.71 | 0.83 | 0.39 |
| | **Random forest** | **BoW** | **0.46** | **0.87** | **0.74** | **0.57** |
| | Naïve Bayes | TF-IDF | 0.22 | 0.62 | 0.89 | 0.35 |
| | Logistic regression | TF-IDF | 0.34 | 0.79 | 0.84 | 0.48 |
| | SVM | TF-IDF | 0.30 | 0.75 | 0.85 | 0.44 |
| | SGD | TF-IDF | 0.22 | 0.64 | 0.83 | 0.34 |
| | **Random forest** | **TF-IDF** | **0.48** | **0.88** | **0.68** | **0.57** |

Коли ми використовували техніку недостатньої вибірки та видаляли екземпляри з мажоритарного класу, оцінка моделей класифікаторів була дуже низькою порівняно з технікою надмірної вибірки. SGD з TF-IDF і наївний Баєс з TF-IDF і BoW вийшли найгіршими з результатами F1 34, 35 і 37% відповідно. З таблиці 4.1 видно, що лише класифікатор випадкових лісів отримав оцінку F1 більше ніж 50%, на відміну від інших класифікаторів, коли застосовувався алгоритм недостатньої вибірки, хоча результати метрики точності були дуже поганими.

На рисунку 4.1 показано порівняння класифікаторів, що використовують різні методи виділення ознак (BoW і TF-IDF) на основі показника F1 (Таблиця 4.1).

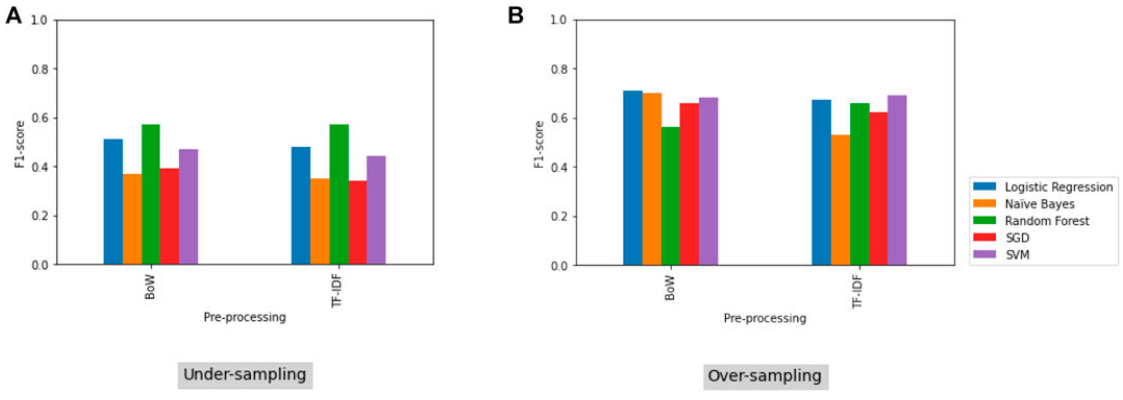

Рисунок 4.1 – Порівняння класифікаторів відповідно до методів виділення ознак

На основі аналізу, який ми виконали в аналізі URL-адрес, було виділено позитивний вплив на оцінку F1 і показники запам'ятовування (рисунок 4.2) у деяких класифікаторах ML, після включення найбільш релевантних функцій, витягнутих з URL-адрес. Як показано в таблиці 4.2, реалізація нових функцій, витягнутих з URL-адрес, успішно допомагала класифікаторам, покращуючи їх продуктивність.

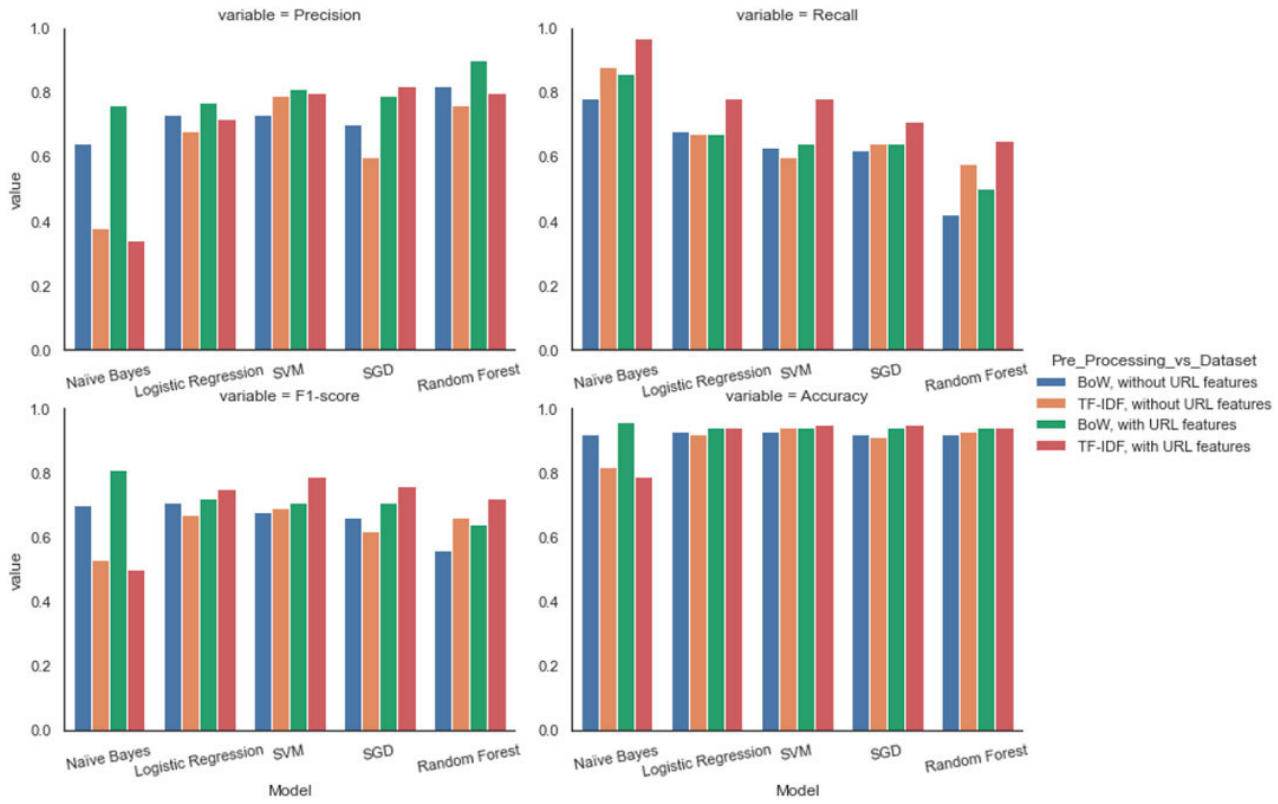


Рисунок 4.2 – Оцінка класифікаторів з використанням URL-аналізу

Таблиця 4.2 – Вплив URL-аналізу на точність класифікаторів

| Re-sampling | Classifier | Pre-processing | Precision | Accuracy | Recall | $F_1$ score |
|---|---|---|---|---|---|---|
| | **Naïve Bayes** | **BoW** | **0.76** | **0.96** | **0.86** | **0.81** |
| | Logistic regression | BoW | 0.77 | 0.94 | 0.67 | 0.72 |
| | SVM | BoW | 0.81 | 0.94 | 0.64 | 0.71 |
| | SGD | BoW | 0.79 | 0.94 | 0.64 | 0.71 |
| Over | Random forest | BoW | 0.90 | 0.94 | 0.50 | 0.64 |
| | Naïve Bayes | TF-IDF | 0.34 | 0.79 | 0.97 | 0.50 |
| | Logistic regression | TF-IDF | 0.72 | 0.94 | 0.78 | 0.75 |
| | **SVM** | **TF-IDF** | **0.80** | **0.95** | **0.78** | **0.79** |
| | SGD | TF-IDF | 0.82 | 0.95 | 0.71 | 0.76 |
| | Random forest | TF-IDF | 0.80 | 0.94 | 0.65 | 0.72 |

Результати підтверджують ефективність запровадження функцій URL-адреси зі значеннями приблизно вище 0,70 для двох типів попередньої обробки. До вибору функції URL найвищий бал F1 був 0,71.

## 4.2 Збалансованість класів в бінарній класифікації

У проблемах бінарної класифікації дисбаланс класів є відкритою проблемою, оскільки набори даних реальних слів зазвичай перекошені. Одне з питань включає визначення найбільш підходящих показників для оцінки ефективності моделі. Оцінка F1, визначена як середнє гармонійне значення точності та запам’ятовування (показники продуктивності), зазвичай використовується для вимірювання рівня дисбалансу. Наші дані мали значно високий рівень дисбалансу (клас більшості, тобто реальні новини, становив приблизно 90% нашого набору даних, а клас меншості, тобто фейкові новини, становив лише 10% набору даних). Способом вирішення та пом’якшення проблеми дисбалансу класів була повторна вибірка даних, яка складається з надмірної або недостатньої вибірки набору даних.

Надмірна вибірка набору даних заснована на перебалансуванні розподілів шляхом доповнення штучно створених екземплярів другорядного класу (тобто фейкових новин). Навпаки, метод недостатньої вибірки заснований на перебалансуванні розподілів шляхом видалення екземплярів класу більшості (тобто реальних новин). Через недостатню вибірку мажоритарного класу нам довелося зменшити розмір вибірки, що в результаті виявилося занадто малим для навчальних моделей, що спричинило низьку продуктивність. За рахунок надмірної вибірки даних ми натомість помітили кращі результати з точки зору як запам’ятовування, так і показників F1, що підвищило продуктивність моделі.

Ми порівняли моделі, засновані на популярних представленнях функцій, таких як BoW і TF-IDF. Після надмірної вибірки даних показники оцінки

повернули результати з оцінками F1 понад 70% як для логістичної регресії, так і для наївних класифікаторів Байєса з BoW.

Щоб покращити результати, ми вирішили також зосередитися на джерелах новин, досліджуючи та вибираючи ознаки URL-адрес, які продемонстрували сильний вплив у різних дослідженнях.

Насправді, як і фішингові атаки (наприклад, підозрілі листи електронної пошти або шкідливі посилання), фейкові новини продовжують викликати головне занепокоєння, оскільки вони все ще поширюються Інтернетом і продовжуватимуть поширюватися, доки всі не зрозуміють, як їх помітити.

З точки зору проблем, двома основними з них є дисбаланс класів реальних даних і обмежена доступність високоякісного міченого набору даних. Використання моделей класифікації ML для виявлення фейкових новин все ще виглядає більш складним у реалістичних ситуаціях, особливо у веб-пошукових системах, де збирається інформація про метадані з тисяч веб-сайтів.

Крім того, як і фішингові атаки, люди (або боти), які пишуть фейкові новини та оманливий вміст, постійно шукають нові та креативні способи обдурити користувачів, щоб вони повірили, що їхні історії пов’язані з надійними джерелами. Це змушує постійно оновлювати моделі, оскільки фейкові новини стають дедалі складнішими і їх важко помітити. Крім того, вміст, що вводить в оману, дуже різниться і змінюється з часом; тому важливо досліджувати нові функції.

# 5 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

## 5.1 Охорона праці

У магістерській роботі досліджено виявлення і запобігання масової дезінформації використовуючи мову програмування Python та фреймворки Scikit-learn та TensorFlow. Обов'язковим елементом дослідження є визначення та аналіз вимог з охорони праці і техніки безпеки при розробці програмного засобу і проведенні експериментальних досліджень, що супроводжується використанням комп'ютерної техніки. Дотримання норм і правил охорони праці є важливим аспектом у контексті дотримання норм організації робочого місця, забезпечення комфортних та зручних умов праці осіб, які беруть участь у процесі, а це вимагає дослідження та дотримання вимог з охорони праці.

В Україні розроблено й діють ряд нормативних документів, які визначають вимоги і правила щодо використання комп'ютерної техніки, приміщень з екранними пристроями та ін. Основним нормативним документом при використанні комп'ютерної техніки є НПАОП 0.00-7.15-18 «Вимоги щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями». Він регламентує, що приміщення для експлуатації комп'ютерної техніки повинно розміщуватися в північній або північно-східній частині будівлі. Площа одного робочого місця повинна становити щонайменше 6 м$^2$, об'єм — щонайменше 20 м$^3$, відстань між робочими столами — щонайменше 2,5 м у ряду і 1,2 м між рядами. Стіни приміщень потрібно фарбувати у пастельні тони з коефіцієнтом відбиття 0,5-0,6 [26].

З метою зменшення напруження очей потрібно, щоб відстань між краями сусідніх точок зображення на моніторі не перевищувала гранично оптимальний розмір літеро-цифрових знаків – 16-20, складних знаків – 35-40. Оптимальні співвідношення параметрів літер і цифр такі: ширина знака – 0,75 їх висоти, товщина ліній при зворотному контрасті – 1/6-1/8, відстань між знаками — 0,25-0,5 висоти знака, між словами – 0,75-1, між рядками – 0,5-1 [26].

Для профілактики загальної втоми і особливо зорового аналізатора важливе значення має організація режиму праці та відпочинку. Загальна тривалість робочого дня не повинна перевищувати 8 год. Частота і тривалість перерв залежать від типу та інтенсивності виконуваних робіт. Під час робіт, які виконуються з великим навантаженням, рекомендуються перерви на 10-15 хв. через кожну годину, а при не інтенсивній і монотонній роботі — на 10-15 хв. через кожні дві години. Кількість мікропауз (тривалістю до хвилини) потрібно регулювати індивідуально.

Зміст регламентованих перерв може бути різний: виробнича гімнастика (вправи для очей, гімнастика, спрямована на корекцію вимушеної робочої пози, поліпшення венозного кровообігу, часткову дисфункцію рухової активності), альтернативна допоміжна робота, приймання їжі тощо.

Для того, щоб особи, які займаються проектуванням та оцінюванням якості людино-машинної взаємодії меншою мірою втомлювались і зберігали високий рівень працездатності, потрібно раціонально організовувати їхні робочі місця. Зокрема, робоче місце має відповідати основним антропометричним даним людини. Крісло або стілець на робочому місці повинні мати висоту сидіння 40-50 см від рівня підлоги, а також відповідний кут нахилу спинки.

Монітори потрібно розміщувати на висоті рівня очей (висота від підлоги до нижнього краю екрана має становити 95-100 см) на відстані 60-70 см від оператора (відстань від краю столу — 50-70 см). Кут зору працюючого щодо екрану має дорівнювати 10-20°, але не більше 40°, кут між верхнім краєм монітора і рівнем очей користувача має становити менш як 10°. Найдоцільніше розміщувати екран перпендикулярно до лінії погляду користувача. Кут нахилу екрана по вертикалі має становити 0-30° [26]. З цією метою сучасні монітори комплектують підставкою з поворотним кронштейном, що дає змогу регулювати кут нахилу монітора і горизонтально обертати його навколо вертикальної осі. Висоту екрана від поверхні підлоги регулюють змінюючи висоту робочої поверхні столу. Іноді монітори встановлюють на спеціальні підставки, що уможливлює його переміщення у просторі у вертикальному та горизонтальному напрямах.

У приміщеннях, де виконуються роботи на ПК, повинно бути передбачене природне і загальне штучне освітлення. Робочі місця користувачів потрібно розміщувати так, щоб у поле зору не потрапляли вікна і освітлювальні прилади (монітори потрібно розміщувати під кутом 90-105° до вікон і на відстані 2,5-4 м від стін і віконних прорізів). У поле зору користувача не повинні потрапляти поверхні, що відбивають світло. Покриття столу має бути матовим з коефіцієнтом відбиття 0,25-0,4.

Для штучного освітлення приміщення рекомендується застосовувати світильники матового світла з розсіювачами, а спектральний склад ламп має наближатися до спектру сонячного світла (наприклад, люмінесцентні типу ЛБ). Оптимальна освітленість робочих місць — 400-500 лк.

У разі ураження електричним струмом необхідно терміново звільнити потерпілого від дії електричного струму (через відключення електроживлення в кімнаті, загального електроживлення на розподільчому щиті або іншим способом). Викликати швидку медичну допомогу (подзвонивши за міським телефоном 103). Надати першу медичну допомогу потерпілому, враховуючи наступне:

- якщо потерпілий знепритомнів, але дихає, його необхідно рівно і зручно вкласти, розстебнути одяг, створити приплив свіжого повітря і забезпечити повний спокій;
- при відсутності ознак життя до прибуття лікарів потерпілому необхідно робити штучне дихання.

При дослідженні виявлення і запобігання масової дезінформації використовуючи мову програмування Python та фреймворки Scikit-learn та TensorFlow було дотримано всіх вищенаведених вимог нормативних документів щодо охорони праці і техніки безпеки при експлуатації комп’ютерної техніки.

## 5.2 Безпека в надзвичайних ситуаціях

Планування заходів цивільного захисту передбачає управління надзвичайними ситуаціями. Для забезпечення безпеки людини в НС управління об’єктами повинно включати здійснення 3-х стратегій:

- запобігання причин виникнення;
- запобігання самих НС;
- пом’якшення, максимальне ослаблення наслідків НС.

Стратегія запобігання причин виникнення НС передбачає недопущення таких дій чи процесів, які несуть загрозу населенню. Дана стратегія здійснюється або відмовою від будівництва небезпечних об'єктів, або знищенням чи перепрофілюванням виробництв – джерел підвищеної небезпеки.

Друга стратегія — запобігання самих НС – передбачає недопущення виходу небезпечного процесу з-під контролю шляхом використання надійних аварійних систем, сигналізації, автоматики й інших заходів з підвищення надійності і стійкості роботи підприємств, а також шляхом заходів превентивної евакуації тощо.

Третя стратегія — пом'якшення, максимальне ослаблення наслідків НС – передбачає орієнтацію на ослаблення, локалізацію наслідків НС. Ця стратегія має пріоритет у керуванні стихійними лихами і ситуаціями «комбінованого» типу.

У практиці управління найбільший ефект дає спільне використання всіх трьох стратегій, особливо при промислових аваріях. У НС, викликаних стихійними лихами, пріоритет надається другій і третій стратегіям. Для реалізації кожної зі стратегій управління необхідно розробляти і приймати комплекс превентивних та оперативних заходів.

Превентивні заходи:

- аналіз і встановлення зовнішніх та внутрішніх причин, які ведуть до НС;
- прогнозування осередків ураження, втрат і збитків на підприємстві;

- заходи з підвищення стійкості;
- обґрунтування сил і засобів для проведення рятувальних та інших невідкладних робіт;
- навчання формувань і робітників діям у НС;
- підготовка надійного командного пункту управління.

Оперативні заходи:

- оповіщення про НС;
- проведення всіх видів розвідки й оцінка обстановки;
- проведення екстрених захисних заходів (укриття в ЗС, евакуація, використання ЗІЗ);
- використання сил постійної готовності для локалізації НС;
- надання першої медичної допомоги;
- нарощування сил і засобів за рахунок залучення формувань підвищеної готовності;
- забезпечення життєдіяльності потерпілих;
- введення аварійно-відновлювальних робіт.

При виникненні НС організовується надзвичайне управління, яке складається з чотирьох стадій ліквідації її наслідків.

Стадія вжиття екстрених заходів. Мета – задіяти механізм надзвичайного управління і вчасно зреагувати на НС. Основні завдання початкової стадії: встановлення факту НС, попередня оцінка обстановки в зоні лиха і масштабів наслідків, мобілізація і визначення оперативних завдань органам надзвичайного управління, віддача розпоряджень на залучення мобільних сил пожежної охорони, швидкої медичної допомоги, охорони громадського порядку й інших служб для допомоги потерпілим, сприяння місцевим органам влади в організації рятувальних робіт і локалізації зони НС власними силами; інформування населення та вищестоящих органів управління про НС і вжиті заходи. Тривалість початкової стадії – 1-10 годин.

Стадія оволодіння ситуацією й організації механізму надзвичайного управління у зоні НС. Завдання: детально оцінити обстановку, терміново прийняти обґрунтоване рішення й уточнити план ліквідації наслідків НС;

розрахувати необхідні сили і засоби, ресурси для всього комплексу робіт у зоні лиха, організувати чітку взаємодію всіх залучених сил і аварійних служб. Тривалість 2-ї стадії – від кількох годин до кількох діб.

Основна і визначальна стадія. Мета – перебороти надзвичайний характер ситуації: відновити безпеку населення в зоні НС, ліквідувати загрозу життю і здоров'ю всім потерпілим, створити мінімально необхідні умови для життєдіяльності населення, що залишилося. Завдання: розгортання в найкоротший термін рятувальних робіт на всіх постраждалих об'єктах зони НС, надання допомоги потерпілим для захисту їхнього життя, здоров'я і підтримка життєздатності в екстремальних умовах; евакуація потерпілих із зони НС та їх життєзабезпечення; термінове проведення аварійно-відновлювальних робіт на системах водо-, тепло-, газо-, електропостачання і зв'язку в зоні НС. Тривалість– кілька діб – кілька тижнів.

Стадія відновлення, тобто економічна, соціальна, культурна екологічна реабілітація зони НС. Органи надзвичайного управління вичерпали свою роль і передають функції постійної дії місцевим органам управління. Розробляється спеціальна програма з черговістю комплексу заходів для реабілітації зони НС.

Отже, планування заходів цивільного захисту на об'єкті у випадку надзвичайних ситуацій безпосередньо є одним з найважливіших процесів для коректної роботи об’єкта. Дотримання визначених правил дозволить здійснити захист працівників об’єкту та майна від надзвичайних ситуацій шляхом запобігання таким ситуаціям, ліквідації їх наслідків і надання допомоги постраждалим.

# ВИСНОВОК

У цьому дослідженні ми проаналізували інформацію метаданих, отриману з веб-пошукових систем, після подання конкретних пошукових запитів, пов'язаних зі спалахом COVID-19, імітуючи звичайну діяльність користувача. Використовуючи як текстові властивості даних, так і URL-адреси, ми навчили різні алгоритми машинного навчання з методами попередньої обробки, такими як пакет слів і TF-IDF. Щоб впоратися з дисбалансом класів через реальні дані, ми застосували методи повторної вибірки, тобто надмірну вибірку фейкових новин і недостатню вибірку реальних новин. У той час як метод надмірної вибірки дозволив нам отримати задовільні результати, метод недостатньої вибірки не зміг підвищити продуктивність моделі, показуючи дуже погані результати через малий розмір вибірки. Хоча новини мають деякі специфічні текстові властивості, які можна використовувати для їх класифікації як підроблених або справжніх, коли ми дивимося на результати пошуку (заголовки, фрагменти та посилання), можна використовувати деяку додаткову попередню обробку, щоб отримати деякі особливі додаткові функції для підробки. Хоча текстові функції пов'язані з новинним вмістом, зібраними як із заголовків, так і з фрагментів, функції URL-адреси базуються на вихідних веб-сайтах, які повертаються як результати пошуку на WSE.

Хоча більшість попередніх досліджень зосереджувалися на виявленні фейкових новин у соціальних мережах, спираючись на дані, які можна отримати безпосередньо з тексту (наприклад, твіти) та використання URL-адрес для підвищення достовірності джерела, запропонований нами підхід йде далі й аналізує особливості URL-адрес саме джерело інформації. Ми дійсно вважаємо, що аналіз шаблону URL-адреси за допомогою методів виявлення фішингу може підвищити здатність алгоритмів ML виявляти та пом'якшувати поширення фейкових новин у всесвітній мережі. Перевірка джерела дійсно є однією з найпоширеніших порад, які веб-сайти з перевірки фактів дають онлайн-читачам. Результати цього дослідження свідчать про те, що інформація про URL-адреси, отримана за допомогою фішингових методів (наприклад, кількість

цифр, кількість крапок і довжина URL-адреси), може дати вказівки дослідникам щодо низки потенційно корисних функцій, які в майбутньому можуть стати фейковими новинами. Алгоритми виявлення можуть мати або розроблятися для того, щоб отримати подальшу цінну інформацію на веб-сайтах, що містять переважно неправдивий вміст, і покращити продуктивність моделі.

Однак аналіз фейкових новин, які поширюються в Інтернеті, може мати потенційні обмеження через пошукову оптимізацію. У цьому дослідженні ми запропонували можливе рішення для її вирішення. Насправді, хоча результати пошукової системи можуть бути налаштовані на основі місцезнаходження користувача в мережі та історії пошуку користувача, щоб зменшити упередження через попередній пошук, було б корисно змінити налаштування налаштувань, видалити кеш, файли cookie та пошук. історію або використовуйте анонімні/приватні вікна. Крім того, використання проксі-серверів (або VPN) може дозволити здійснювати пошук запитів, які не залежать від географічного розташування.

З точки зору майбутніх досліджень виявлення фейкових новин, ми вважаємо, що методи, які зазвичай використовуються для виявлення шкідливих URL-адрес, також слід розглянути для виявлення фейкових новин: це означатиме створення класифікаторів на основі не лише традиційних лексичних і семантичних особливостей текстів, а й лексичних та ознак URL-адреси.

# ПЕРЕЛІК ПОСИЛАНЬ


1. IBM SpectrumStorage [Електронний ресурс] – Режим доступу до ресурсу: https://www.techrepublic.com/resource-library/whitepapers/ibm-spectrumstorage-making-the-impossible-possible/.
2. Six types of misinformation [Електронний ресурс] – Режим доступу до ресурсу: https://www.cjr.org/tow_center/6_types_election_fake_news.php.
3. Many People Say Made-Up News Is a Critical Problem That Needs To Be Fixed [Електронний ресурс] – Режим доступу до ресурсу: https://www.pewresearch.org/journalism/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/.
4. Buzzword Or Real Threat? Fake News Is More Dangerous Than You Think. / [Sebastiaan van der Lans]. – 2019
5. Fake News Detection via NLP is Vulnerable to Adversarial Attacks. / [Zhixuan Zhou, Huankang Guan, Meghana Moorthy Bhat and Justin Hsu] // – 2019
6. Spam filtering in twitter using sender-receiver relationship. / [J. Song, S. Lee, and J. Kim] // – 2018
7. Twitter Spam Detection based on Deep Learning. / [Tingmin Wu, Shigang Liu, Jun Zhang and Yang Xiang] // – 2017
8. Weighted and Probabilistic Context-Free Grammars Are Equally Expressive. / [Noah A. Smith, Mark Johnson] // – 2017
9. New Knowledge [Електронний ресурс] – Режим доступу до ресурсу: https://www.newknowledge.com.
10. Big Data and quality data for fake news and misinformation detection. / [Fatemeh Torabi Asr, Maite Taboad] // – 2019
11. The Partnership Press: Lessons for Platform-Publisher Collaborations as Facebook and News Outlets Team to Fight Misinformation. / [Mike Ananny] // – 2018
12. Computational Fact Checking from Knowledge Networks. / [Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M. Rocha, Johan Bollen, Filippo Menczer, Alessandro Flammini] // – 2017

13. ClaimRank: Detecting Check-Worthy Claims in Arabic and English. / [Israa Jaradat, Pepa Gencheva, Alberto Barron-Cedeno, Lluis Marquez, Preslav Nakov] // – 2018
14. A Stylometric Inquiry into Hyperpartisan and Fake News. / [Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, Benno Stein] // – 2017
15. Syntactic Stylometry for Deception Detection. / [Song Feng, Ritwik Banerjee, Yejin Choi] // – 2012
16. CSI: A Hybrid Deep Model for Fake News Detection. / [Natali Ruchansky, Sungyong Seo, Yan Liu] // – 2017
17. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. / [Svitlana Volkova, Kyle Shaffer, Jin Yea Jang and Nathan Hodas] // – 2017
18. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. / [Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, Yejin Choi] // – 2017
19. Claas Relotius Reporter Forgery Scandal [Електронний ресурс] – Режим доступу до ресурсу: http://www.spiegel.de/international/zeitgeist/claas-relotius-reporter-forgery-scandal-a-1244755.html
20. Hyperpartisan Facebook Pages Are Publishing False And Misleading Information At An Alarming Rate [Електронний ресурс] – Режим доступу до ресурсу: https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis.
21. Neural Network Methods in Natural Language Processing. / [Yoav Goldberg] // – 2017
22. Морфологічний аналізатор pymorphy2 [Електронний ресурс] – Режим доступу до ресурсу: https://pymorphy2.readthedocs.io/en/stable/
23. Морфологічний аналізатор Dostoevsky [Електронний ресурс] – Режим доступу до ресурсу: https://vc.ru/ml/144551-dostoevsky-biblioteka-analiza-nastroeniy-dlya-russkogo-yazyka.
24. Фреймворк Angular [Електронний ресурс] – Режим доступу до ресурсу: https://angular.io/guide/what-is-angular.

25. An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data. / [Xie W, Liang G, Dong Z, Tan B, Zhang B.] // – 2019
26. Вимоги щодо безпеки та захисту здоров’я працівників під час роботи з екранними пристроями [Електронний ресурс] – Режим доступу до ресурсу: https://zakon.rada.gov.ua/laws/show/z0508-18#n14.

# ДОДАТКИ