Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Програмної інженерії
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

другий (магістерський)
(рівень вищої освіти)

Дослідження методів лінгвістичного аналізу відгуків користувачів інтернет-форумів
(тема)

Виконав: студент 2 курсу, групи ПЗСм-18-1
(курс, назва групи)

Давидов О.П.
(прізвище, ініціали)

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

Освітньо-професійної програми
(тип програми)

Програмне забезпечення систем
(повна назва освітньої програми)

Керівник: к.т.н. доцент Голян В.В.
(посада, прізвище, ініціали)

Допускається до захисту
Зав. кафедри, проф. _____ Дудар З.В.
(підпис)

2019 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Програмної інженерії

Рівень вищої освіти другий (магістерський)

Спеціальність 121-Інженерія програмного забезпечення

Тип програми освітньо-професійна програма

Освітня програма Програмне забезпечення систем

ЗАТВЕРДЖУЮ:

Зав. кафедри _____
(підпис)

«_____» _____ 2019 р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові Давидову Олександру Павловичу
(прізвище, ім'я, по батькові)

1. Тема роботи: *Дослідження методів лінгвістичного аналізу відгуків користувачів інтернет- форумів.*
затверджена наказом університету від «___» _____ 2019 р. № _____
2. Термін подання студентом роботи до екзаменаційної комісії _____ 2019 р.
3. Вихідні дані до роботи: *критерії пошуку спаму, прототип системи для пошуку спаму, пояснювальна записка. Використовувати ОС Windows.*
4. Перелік питань, що потрібно опрацювати в роботі: *мета роботи, аналіз проблемної галузі і постановка задачі, огляд різновидів спаму в мережі Інтернет та їх класифікації, огляд методів пошуку спаму в мережі Інтернет, огляд методів лінгвістичного аналізу, огляд критеріїв для виявлення спаму, перевірка можливості використання отриманих критеріїв спаму для пошуку спаму.*

5. Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		Підпис	Дата
Спецчастина	доц. Голян В.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Аналіз предметної галузі	20 вересня 2019р.	
2.	Огляд існуючих методів	7 жовтня 2019р.	
3.	Дослідження методів лінгвістичного аналізу відгуків користувачів інтернет-форумів	18 листопада 2019р.	
4.	Підготовка пояснювальної записки	27 листопада 2019р.	
5.	Підготовка презентації та доповіді	5 грудня 2019р.	
6.	Попередній захист	10 грудня 2019р.	
7.	Нормоконтроль, рецензування	17 грудня 2019р.	
8.	Занесення диплома в електронний архів	17 грудня 2019р.	
9.	Допуск до захисту у зав. Кафедри	17грудня 2019р.	

Дата видачі завдання «___» _____2019 р.

Студент _____ Давидов О.П.
(підпис)

Керівник роботи _____ к.т.н. доц. Голян В.В.
(підпис)

РЕФЕРАТ / ABSTRACT

Атестаційна робота містить: 73 с., 13 рис., 7 табл. 12 формул, 4 додатки, 22 джерела, графічна частина 16 аркушів.

ДОПИС, ІНТЕРНЕТ, КОМЕНТАР, КРИТЕРІЇ, ЛІНГВІСТИЧНИЙ АНАЛІЗ, МЕТОД ОПОРНИХ ВЕКТОРІВ , СПАМ.

Об’єкт дослідження – методи лінгвістичного аналізу та критерії спаму, отримані на їх основі.

Метою роботи дослідження є встановлення універсальних критеріїв спаму на основі методів лінгвістичного аналізу та перевірка їх ефективності, з використанням методу опорних векторів.

Дослідження базується на використанні методів лінгвістичного аналізу. Розробка прототипу базується на використанні мови програмування Python, бібліотек NLTK, для обробки тексту, та profanity-check, для пошуку нецензурної лексики. Серед математичних методів були використані: нормалізація, середнє значення та мода, стандартне відхилення.

У результаті дослідження було проаналізовано предметну галузь, розглянуто роботи з ідентифікації спаму попередників, розглянуто різні критерії спаму та встановлено точність прототипу спам аналізатору на основі цих критеріїв та методу опорних векторів з використанням датасету Мішне.

POST, INTERNET, COMMENTARY, CRITERION, LINGUISTIC ANALYSIS, SUPPORT VECTOR MACHINE, SPAM.

The object of research is the methods of linguistic analysis and criteria based on them.

The purpose of the study is to establish universal criteria for spam based on linguistic analysis methods and to verify their effectiveness using the support vector machine.

The study is based on the use of linguistic analysis methods. Prototype development is based on the use of Python programming language, NLTK libraries, for word processing, and profanity-check, for searching obscene language. Among the mathematical methods were used: normalization, mean and mode, standard deviation.

The study analyzed the subject area, reviewed the work of identifying spam precursors, examined different spam criteria, and established the accuracy of a prototype spam analyzer based on these criteria and the method of reference vectors using the Mishne dataset.

ЗМІСТ

ВСТУП

На даний період часу популярність вебу обумовлена популярністю веб додатків, що дозоляють розповсюджувати інформацію, медіа та різні види іншого контенту. Такі додатки зазвичай виступають також платформами, що забезпечують можливість людям висловлювати власні думки, а також отримувати у розпорядження інформацію про думки сотень та тисяч інших користувачів мережі Інтернет. Зазвичай одиницями, що являють собою думку якоїсь конкретної особи, виступають коментарі. Такий метод обміном інформації, набув надзвичайної популярності завдяки таким платформам як YouTube, Reddit, Twitter, Facebook, Instagram та ін.

Проте не завжди люди бажають залишити коментар, щоб їх думку побачили та оцінили, це також і золота жила для людей, що хочуть розповсюджувати спам. Завданням спамерів є заплутати користувача і таким способом спровокувати його перейти до іншого веб сайту, аби підняти його трафік та рейтинги. Також зустрічаються випадки, коли засобами образ та нецензурної лексики спамери намагаються змусити людину написати щось у відповідь, що призводить до збільшення рейтингів тієї чи іншої статті під якою розверзалися словесні баталії.

Для того щоб повністю оцінити масштаб проблеми наведемо статистичні данні: лише за третій квартал 2019 року платформа YouTube була вимушена видалити близько 426 мільйонів коментарів [21]. За словами самої компанії, це лише мала частка серед мільярдів коментарів, що були розміщені за цей період.

Крім того, що неможливо не погодитись з тим, що спам є досить дратуючим сам по собі, він також може нести в собі загрозу для користувача та його комп’ютера. Посилання, що досить часто містяться в спам-коментарях, можуть переадресовувати користувача не лише на рекламні ресурси, а й на ресурси сумнівного характеру, які можуть розповсюджувати шкідливе програмне забезпечення.

Також слід звернути увагу на те, що в другій чверті 2019 року спамери почали активно використовувати хмарні сервіси для того, щоб маскувати свої посилання [22]. Причиною використання сервісів Google та інших відомих компаній є те, що посилання на легальний домен є менш підозрілими.

Актуальною проблема спаму стала ще на момент 90-х років минулого століття. За цей час була проведена велика кількість досліджень. Більшість ефективних методів були засновані на методах лінгвістичного аналізу. Для розробки цих методів було проаналізовано склад спам-коментарів, їх властивості. На основі цього аналізу були виділені основні відмінності за якими можливо визначити чи є коментар спамом. Проте всі раніше застосовувані методи спираються на складні ключові критерії, що працюють досить ефективно лише на наборах даних, що схожі з використаними для їх навчання.

Метою атестаційної роботи є покращення вже існуючих алгоритмів розпізнавання спаму в сфері інтернет-форумів та блогів та пошук простих критеріїв спаму на основі методів лінгвістичного аналізу.

Завданням атестаційної роботи є дослідження методів лінгвістичного аналізу відгуків користувачів інтернет-форумів та блогів. Завдяки методам лінгвістичного аналізу планується виявити особливості та відмінності між звичайними та спам-коментарями. На основі цих особливостей планується виділити критерії для аналізу коментарів. Після чого буде проведене навчання алгоритму для автоматичного виявлення спаму. Також буде проведене порівняння результатів отриманого алгоритму з результатами попередніх розробок.

1 АНАЛІЗ ПРОБЛЕМНОЇ ГАЛУЗІ

1.1 Загальний огляд проблеми

Соціальні мережі дозволяють користувачеві проводити аналіз відгуків того чи іншого товару на основі його опису та коментарів користувачів, покупців, спеціалістів чи просто зацікавлених осіб. Проте не завжди коментарі містять необхідну чи корисну інформацію. Дуже часто коментарі виступають у якості спам-повідомлень для користувачів сайтів. Основна задача таких спам матеріалів є зацікавлення користувачів та рекламування їм сторонніх ресурсів чи товарів.

На найбільш популярних сайтах існує спосіб боротьби з такими коментарями на основі відгуків. Якщо коментар має низький рейтинг – його видаляють. Таким підходом користується YouTube, він видаляє такі коментарі [6, 9]. Схожий метод працює і на сайті Cybersport.ru, проте на цьому сайті вони не видаляються, а просто замилюються для того, щоб їх було складніше помітити.

Вищеописаний підхід не є високопродуктивним, адже з моменту написання коментаря й до моменту його видалення може пройти досить великий проміжок часу (або він взагалі не отримає необхідної кількості голосів, необхідних для видалення).

Проте у спам-коментарів є чіткі властивості, що простежуються від одного до іншого. До них можна віднести: уривчастість тексту, вульгарну або нецензурну лексику в повідомленні, тема коментаря не має нічого спільного з темою, що обговорюється. Також використовують візуальну подачу коментаря, коли його виділяють відступами. І одним з найбільш виразних символів спам-коментарів є наявність посилань на веб ресурси.

Знаючи про ці властивості спаму можливо побудувати систему, яка, використовуючи методи лінгвістичного аналізу, обробки природної мови та машинного навчання, буде відсіювати спам-коментарі [8].

1.2 Актуальність проблеми

Проблема спаму надзвичайно актуальна в наш час. Розповсюдженість блогів, медіа-майданчиків, соціальних мереж та інших платформ, на яких є можливість коментування призвела до буму спаму в коментарях, щоб усвідомити масштаби проблеми приведемо статистику платформи YouTube.

Згідно зі статистикою, що надає нам компанія Google, кількість видалених коментарів за період з липня по вересень 2019 року становить 516 887 894. З них 82% (426 192 696) було ідентифіковано як спам повідомлення [21]. На рисунку 1.1 наведено діаграму видалення коментарів за типами порушень.

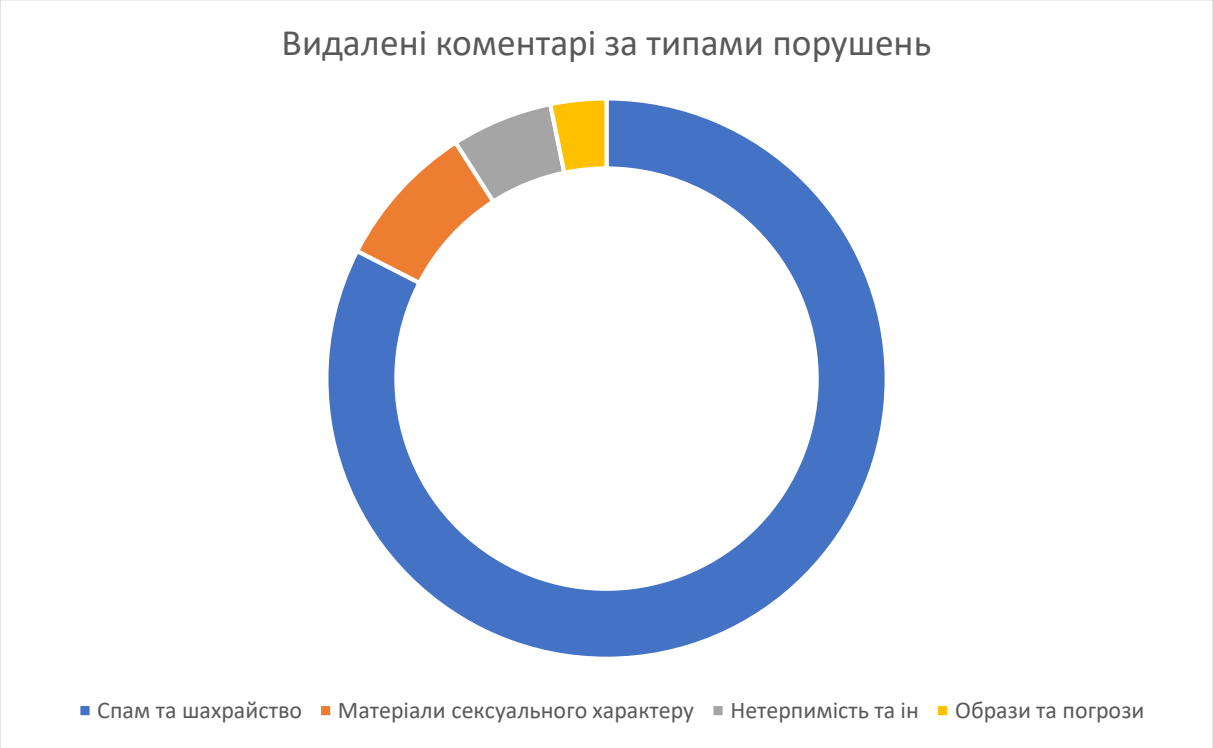

Рисунок 1.1 – Видалені коментарі за типами порушень

У той же час, за той самий період 2018 року було видалено лише 166 370 039 коментарів. А це означає, що за рік кількість видалених коментарів зросла в 3 рази. Загальна кількість видалених каналів становить 3 315 189, з них 91,1% (3 020 702) за спам. Загальна статистика по видаленням каналів наведена в таблиці 1.1.

Таблиця 1.1 – Статистика видалення каналів на платформі YouTube

	Проміжок часу				
	07.2018-09.2018	10.2018-12.2018	01.2019-03.2019	04.2019-06.2019	07.2019-09.2019
Кількість видалених каналів	1 667 587	2 398 961	2 828 221	4 069 349	3 315 189
Відсоток видалених за спам	79,6% (лише за 09.2018)	81,7%	85,1%	90,3%	91,1%
Кількість видалених за спам	468 458 (лише за 09.2018)	1 960 498	2 407 420	3 676 012	3 020 702

Згідно з наведеними вище даними, можна зробити висновок, що кількість спаму в мережі Інтернет не просто не спадає, а й з кожним роком зростає, тому ця проблема є надзвичайно актуальною на період 2019 року. Ріст проявляється у збільшенні не тільки загальної кількості коментарів та каналів, а й у збільшенні відсотка видалень через спам та шахрайство порівняно з іншими причинами видалень.

1.3 Інтернет-спам

Інтернет-спам – поняття, яке впродовж останніх 20-ти років розширялось, доповнювалось та розвивалось завдяки появі різних його видів та підвидів. Зародженням інтернет-спаму ми маємо бути вдячні електронній пошті та першому з усіх видів спаму в мережі Інтернет – електронним спам-листам.

В таблиці 1.2 відображена загальна класифікація Інтернет-спаму за його підтипами та його цілями.

Таблиця 1.2 – Типи Інтернет-спаму.

Тип спаму	Підтип спаму	Направленість
Email-спам		Безпосередньо на користувача, фішинг
Спам-повідомлення		Безпосередньо на користувача
Веб-спам	Маніпулювання алгоритмами пошуку, спам-назви, спам в метатегах, URL-маскування	Безпосередньо на користувача, алгоритми пошуку
Блог-спам	Спам-блоги	Безпосередньо на користувача, алгоритми пошуку
	Спам-коментарі	Безпосередньо на користувача, алгоритми пошуку
	Трекбек-спам (Trackback Spam)	

Як видно з таблиці 1.2 спам поділяється на 4 основні категорії. Різниця між цими видами полягає не лише в джерелах його розсилки, а також і в орієнтованості на кінцевого користувача, або на алгоритми пошуку.

Найбільшого масштабу останнім часом набуває блог-спам. Така тенденція обумовлена появою доступних методів створення блогів, великою кількістю ресурсів, на яких доступна функція коментування та високою зацікавленістю аудиторії в цих ресурсах.

Привабливість блог-спаму для зловмисників також полягає в можливості впливати не тільки на користувача блогів, а й на алгоритми пошукових систем, дозволяючи виводити в топ видачі пошукових запитів відверто неякісний контент.

Отже, можна зробити висновок, що найбільш продуктивним у плані задіяні ресурси по відношенню до потенційно охопленої аудиторії є блог-спам.

Також блог-спам є найбільш варіативним в плані цілей типом розповсюдження спаму.

1.4 Спам у сфері блогів та форумів, його види

На відміну від спаму в електронних листах, спам у сфері блогів та форумів розділився на декілька підвидів, кожен з яких має свої особливості та відмінності.

Підвиди спаму у сфері блогів:

– спам блоги – це інтернет-блоги в яких сам допис чи стаття використовується в якості рекламного матеріалу для продуктів чи сервісів. Згідно з дослідженням Google (Google Enterprise, 2009) найбільшого масштабу проблема досягла у 2009 році;

– спам-коментарі – підвид спаму який не має до блогу чи статті жодного відношення. На відміну від спам-блогів, спам-коментарі орієнтовані на всі типи блогів, що дозволяють коментування [3]. Популярний сервіс Akismet дозволяє виявити спам-коментарі з точністю 82% [12];

– трекбек спам (trackback spam) – спам, що заснован на функціях трекбек-пінгу (trackback-ping). Мета спамера – отримати посилання від популярного блогу. Специфікація трекбек-технології не передбачає верифікації, на відміну від коментування безпосередньо на сайті, що дозволяє спамерам вставляти будь-які URL в трекбек-повідомлення, а також маскувати текст повідомлення.

На відміну від інших двох типів, спам-коментарі більш популярні та більш розповсюдженні. Зокрема, їх можна зустріти не тільки в блогах та статтях, а й під будь-яким онлайн медіа-ресурсами, що дозволяють коментування.

1.5 Спам-коментарі

Наразі популярними являють такі види спаму як: коментарі до статей та відео. Це відбувається через те, що люди висловлюють свою позицію в них щиро,

готові її відстоювати, тобто їх досить легко підчепити на гачок дискусії і, маніпулюючи їх відвертістю, змусити перейти на ресурс, що рекламує спамер.

Так як соціальні мережі та блоги є однією з категорій вебу, що розвиваються найбільш швидко і орієнтовані на широку аудиторію в усіх аспектах (починаючи від віку і закінчуючи вподобаннями), вони також є одними з найбільш вразливих до проблем пов'язаних з розповсюдженням спаму.

Зараз найбільш популярними в цій сфері є спам у вигляді коментарів до статей та відео. Для того щоб краще зрозуміти, як працюють спам технології треба розглянути два напрямки, в яких розгортають свої дії спамери.

Першим з них є втягнення в дискусію. Такий підхід характерний тоді, коли спамеру необхідно підняти активність в коментарях під статтею чи відео. В цьому випадку його основною метою буде образити власника матеріалів або власника одного з найпопулярніших за кількістю лайків коментаря, або розпочати в коментарях дискусію на одну із суперечливих тем. Завдяки тому, що в мережі Інтернет тобі гарантована анонімність ти можеш висловлювати свої думки відверто. Саме на це і сподівається спамер, його мета – образити думки якомога більшої кількості користувачів і таким чином отримати більшу кількість образливих коментарів у свою адресу, що веде до збільшення активності на сайті.

Другий підхід – зацікавити користувача та змусити його перейти на сторонній ресурс. В цьому випадку його метою є підняти активність на сайті за допомогою сторонніх ресурсів. Для цього він буде маніпулювати цікавістю користувачів, намагатиметься виділити свій коментар на фоні інших.

Сам процес спаму полягає в розміщені коментарів до публічних блогів, що дозволяють користувачам переходити на сторонні ресурси. Спам використовується для заробітку за рахунок реклами, кліків, встановлення рекламного програмного забезпечення та ураження системи програмним забезпеченням з метою крадіжки персональних даних. Спам-коментарі зазвичай містять посилання на спам-сайти. Також завдяки ним в мережі Інтернет просувається неетичний контент, який, за звичайних умов, мав би низький рейтинг та низькі позиції у видачі в пошукових системах.

Відкритість блогів до коментування робить їх мішенями для зловживання, головною причиною слугує відсутність або неспроможність системи фільтрів відсіяти спам-коментарі. Також, слід відзначити, що спамери уникають блокування та потрапляння до чорних списків за рахунок створення випадкових або використання динамічних IP-адрес. CAPTCHA (Complete Automated Public Turing test to tell Computers and Humans Apart) є популярним засобом боротьби зі спамом (і в деякій мірі досить успішним засобом), проте не вирішує проблему, так як надто проста CAPTCHA досить легко проходиться алгоритмами машинного навчання. І навпаки, занадто складна – викликає проблеми у користувачів інтернет-ресурсу і відбиває будь-яке бажання коментувати.

Іншим популярним підходом є використання фільтру за ключовими словами (фразами). Проте такий підхід вимагає постійного моніторингу та оновлення списку заборонених слів, а також призводить до блокування цілком легітимних коментарів.

1.6 Різниця між спам-коментарями та email-спамом

Email-спам та спам-коментарі досить схожі за деякими показниками, але вони також мають і відмінності. Однією з основних відмінностей є те, що email-спам цілком спрямований на те, щоб людина відвідала той чи інший сайт, чи купила прорекламований продукт. В свою чергу, спам-коментарі можуть орієнтуватися на функції пошукових систем, для того щоб підвищити частоту або позицію на сторінці результатів пошуку. Більш того спам-коментарі охоплюють більшу аудиторію, так як вони доступні для кожного відвідувача блогу, в той час коли email-спам орієнтований на конкретну особу (власника email адреси, на який було відправлено спам-лист). Також, ефект від спам-коментаря миттєвий, адже він починає працювати одразу після публікації. Роботу ж email-спаму майже неможливо відстежити, адже він може і не дістатися до своєї цілі, він може бути

відсіяним завдяки системі фільтрів електронних листів. Інша різниця полягає у тому, що email-спам використовує зображення, а у випадку спам-коментарів HTML тег <*img*> в більшості випадків відфільтровується.

1.7 Існуючі роботи з виявлення спаму

Перші дослідження спаму стосуються саме email-спаму.

Перша роботи з виявлення спаму датована ще 1998-м роком і проведена групою, до якої входили Сахамі, Думаіс, Хекерман та Хорвітц [1]. Вони застосували наївний баєсів класифікатор (Naive Bayes Classifiers), для того щоб класифікувати email-повідомлення.

В 1999-му році Дракер, Ву та Вапнік [13] застосували метод опорних векторів (Support Vector Machines) у боротьбі зі спамом.

В 2001-му році Каррера та Маркез [5] продемонстрували, що AdaBoost (Adaptive Boosting) є більш ефективним за дерево рішень та наївний баєсів класифікатор.

В 2004-му році Джанг, Джу і Йао [11] використали спам-посилання як критерій та порівняли наївний баєсівський класифікатор, метод опорних векторів та LogitBoost.

Всі вищеназвані дослідження були спрямовані на дослідження email-спаму і ці методи виявлення спаму покладались на аналіз вмісту електронних листів.

Наприкінці 1990-х років у зв’язку з ростом популярності мережі Інтернет спам почав розповсюджуватись зі сфери електронного листування на сферу блогів. Саме в цей час дослідники спаму та розробники програмного забезпечення для захисту від нього розвивають роботи в таких напрямках як: веб спам, спам-блоги, спам-коментарі. Різні методи фільтрування веб спаму в цей час використовують як аналіз вмісту так і аналіз посилань для виявлення спам-сторінок в Інтернеті.

В 2005-му році Бечетті, Кастільйо, Донато, Леонарді та Баеза-Єйтс використали дерева рішень для визначення спаму на основі посилань, за допомогою алгоритмів PageRank та TrustRank [14]. В цьому ж році, Дрост та Шеффер застосували метод опорних векторів для класифікації веб-спаму на основі вмісту та посилань [10]. В цей же час, компанія Umbria випустила звіт [15], в якому наголосила, що сфера блогів в Інтернеті зіткнулась з проблемами спаму. Джионджи та Гектор випускають власну класифікацію спаму [16], в якій вони розкривають різні типи спаму, техніки генерації спаму, техніки маскування спаму.

В 2006-му році Нтоулас, Найорк, Манассе та Феттерлі побудували дерево рішень для класифікації веб-спаму на основі його вмісту [17]. Коларі, Джава, Фінін. Оатс та Джоши використали метод опорних векторів на основі посилань, N-gram та bag-of-words критеріїв [18]. Хан, Ахн, Мун, та Джеонг запропонували колаборaційний метод фільтрації, для боротьби зі спам-посиланнями, який засновується на ручній ідентифікації спаму та розповсюдження цієї інформації в мережі пошуку [19]. Мішне розробив мовні моделі для дописів та статей [2], коментарів та сторінок, на які посилаються коментарі, в якому коментарі класифікувались за розбіжностями в мовних моделях.

В 2007 році Кормак, Гомез та Санс провели аналіз фільтрації коротких повідомлень [20]. Вони оцінили різні методи фільтрації на основі вмісту, які використовували такі алгоритми як: наївний баєсів класифікатор; метод опорних векторів; динамічне стиснення Маркова; логістичну регресію, що використовує критерій bag-of-words.

В таблиці 1.3 підсумовано всі вищеописані підходи, та інші підходи до аналізу коротких коментарів.

Таблиця 1.3 – Існуючі підходи до визначення спаму.

Цілі спаму	Існуючі роботи	Критерії виділення спаму	Метод (алгоритм) пошуку спаму	Найкращий результат
Коментарі	Мішне та Кармел (2005)	Bag-of-words	Відстань Кульбака-Лейблера	83% точності

Кінець таблиці 1.3

Цілі спаму	Існуючі роботи	Критерії виділення спаму	Метод (алгоритм) пошуку спаму	Найкращий результат
Статті	Коларі та ін (2006)	N-gram, bag-of-words, tf-idf weighting	Метод опорних векторів	0,9 AUC
Статті	Хан та ін (2007)	Ручний метод, заснований на довірі	Спільний обмін	80% точності
Короткі повідомлення(смс, коментарі)	Кормак та ін (2007)	Bag-of-words, розширений опираючись на позицію в тексті	Динамічне стиснення Маркова, метод опорних векторів, логістична регресія	0,95 AUC
Статті	Ішида (2008)	Ключові слова, посилання	Фільтрація на основі переходів	53% точності
Коментарі	Ромеро, Гарсіа-Вальдес та Аланіс (2010)	Bag-of-words, tf-idf weighting	Наївний баєсів класифікатор, К-найближчих сусідів, нейронні мережі, метод опорних векторів	84,6% точності
Коментарі	Хуан, Джианг, Жанг (2010)	Довжина коментаря, відстань Кульбака-Лейблера, відношення популярних слів	Метод опорних векторів, наївний баєсів класифікатор, дерева рішень	92,86% точності

Більшість з описаних методів визначення спаму використовують початковий набір навчальних даних для створення системи. Як тільки вона побудована, вона може бути використана для визначення подальших спам-коментарів. Цей підхід має як свої сильні сторони, так і слабкі. На схожому до навчального наборі даних точність буде досить висока. Проте спам-коментарі постійно змінюються, тому такій системі буде досить важко пристосуватись до змін. Також, слід зазначити, що не всі критерії підходять для роботи з коментарями. Такі критерії як: bag-of-words та n-grams, дають точні значення на великих документах.

1.8 Лінгвістичний аналіз та його методи

Лінгвістичний аналіз в мовознавстві – це розгляд мовної структури тексту та пошук закономірностей.

Лінгвістичний аналіз використовується в програмуванні для того, щоб отримати значення з тексту. Обробка природної мови цілком і повністю спирається на методи лінгвістичного аналізу.

Методи лінгвістичного аналізу:

– визначення речень: система намагається визначити речення в тексті. Основна проблема полягає в тому, що більшість систем знаходять лише по одному реченню зараз. Такий підхід полегшує виконання, проте ускладнює визначень теми.

– лексичний аналіз (tokenization): на цьому етапі система розбиває речення на слова. Цей метод досить сильно залежить від якості тексту, що надається.

– лематизація та очищення: на цьому етапі всі слова перетворюються на леми, тобто на базову версію слова (наприклад «деревами» перетворюється на «дерево»). Зазвичай лематизація використовують таблиці пошуку, та спеціальні алгоритми що прибирають форму множини та ін. Також на цьому етапі відбувається виправлення помарок та перетворення смайликів на відповідні слова.

– визначення частини мови: на цьому етапі для кожного з токенів (слів) встановлюється частина мови таких як: іменник, прикметник, дієслово та ін.

Вищеописані методи зазвичай передують встановленню значення речення

Лінгвістичний аналіз на основі правил – це підхід, що базується на отриманні результату, що задовольняє певним правилам без розуміння людської мови. Аналіз на основі правил завжди фокусується на єдиній задачі. Прикладом аналізу на основі правил може слугувати пошук пари слів, перше з яких не є дієсловом, а інше – дієслово. Тобто ми використовуємо цей підхід, коли взаємодіємо з певними шаблонами.

1.9 Особливості дослідження

На відміну від досліджень попередників, було прийнято рішення використати низку простих критеріїв, таких як: кількість речень, доля пунктуаційних символів у коментарі та ін. Це зумовлено тим, що, зазвичай, класифікатори використовую досить складні або специфічні критерії для пошуку спаму, тому їх результати цілком залежать від схожості набору даних, що був наданий на класифікацію, та набору даних, що був використаний для навчання алгоритму класифікації.

Використання простих критеріїв, визначених за допомогою методів лінгвістичного аналізу, і постійне навчання нашого алгоритму дозволить нам уникнути ситуації «мутації» спаму (перебудови спам повідомлень під критерії пошуку задля маскування), адже основні принципи його роботи, як і структура завжди залишаються незмінними. В подальшому ці критерії можуть бути використані для аналізу інших типів спаму, наприклад: спам-повідомлень, email-спаму та ін.

1.10 Постановка задачі

Дослідити методи лінгвістичного аналізу та роботи попередників в області спаму, на їх основі сформувати критерії спаму для інтернет-коментарів до дописів. Визначити критерії для спам аналізатора. Враховуючи сформовані критерії спаму та взявши за основу метод опорних векторів побудувати прототип спам-аналізатора. Використовуючи побудований прототип спам-аналізатору, розглянути його показники на прикладі датасету Мішне та Кармел. Визначити показники прототипу спам аналізатора згідно з визначеними раніше критеріями. Порівняти результати з вже існуючими на даний момент роботами.

Виходячи з результатів, зробити висновок про можливість застосування критеріїв спаму та подальші перспективи розвитку роботи.

2 ХАРАКТЕРИСТИКИ СПАМ-КОМЕНТАРІВ

Для того щоб проаналізувати характеристики спаму ми використовуємо збірник створений Мішне та Кармел (2005). Цей збірник містить приблизно 50 статей з 1024 коментарями до них. Всі статті містять суміш з легітимних та спам-коментарів. Загалом 332 коментаря – легальні, інші – спам. В результаті аналізу було виведено критерії спаму, що будуть використані для пошуку спаму.

В якості критеріїв спаму було вибрано:

– кількість пробілів;

– кількість речень;

– кількість посилань;

– доля пунктуаційних символів в загальній кількості символів;

– доля біграм відносно усіх слів;

– доля унікальних слів;

– схожість статті та коментаря;

– вірогідність нецензурної мови.

2.1 Схожість статті та коментаря

Зазвичай спамери використовують скрипти для створення великої кількості спам-коментарів. Однак, в більшості випадків автоматично-згенерований спам не має жодного відношення до теми блогу чи статті. Ми аналізуємо схожість статті та коментаря, і, в результаті, легітимний коментар має включати в себе більшу кількість фраз чи слів, що наявні в статті. Числове значення цієї характеристики розраховується за формулою 2.1:

$$Similarity = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,k}^2}} \quad (2.1)$$

де $w_{i,j}$ – це частота появи слова у блозі чи статті;

$w_{i,k}$ – частота появи слова у коментарі.

На рис 2.1 зображено графік, що показує значення схожості коментаря та статті для легітимних коментарів та спаму.

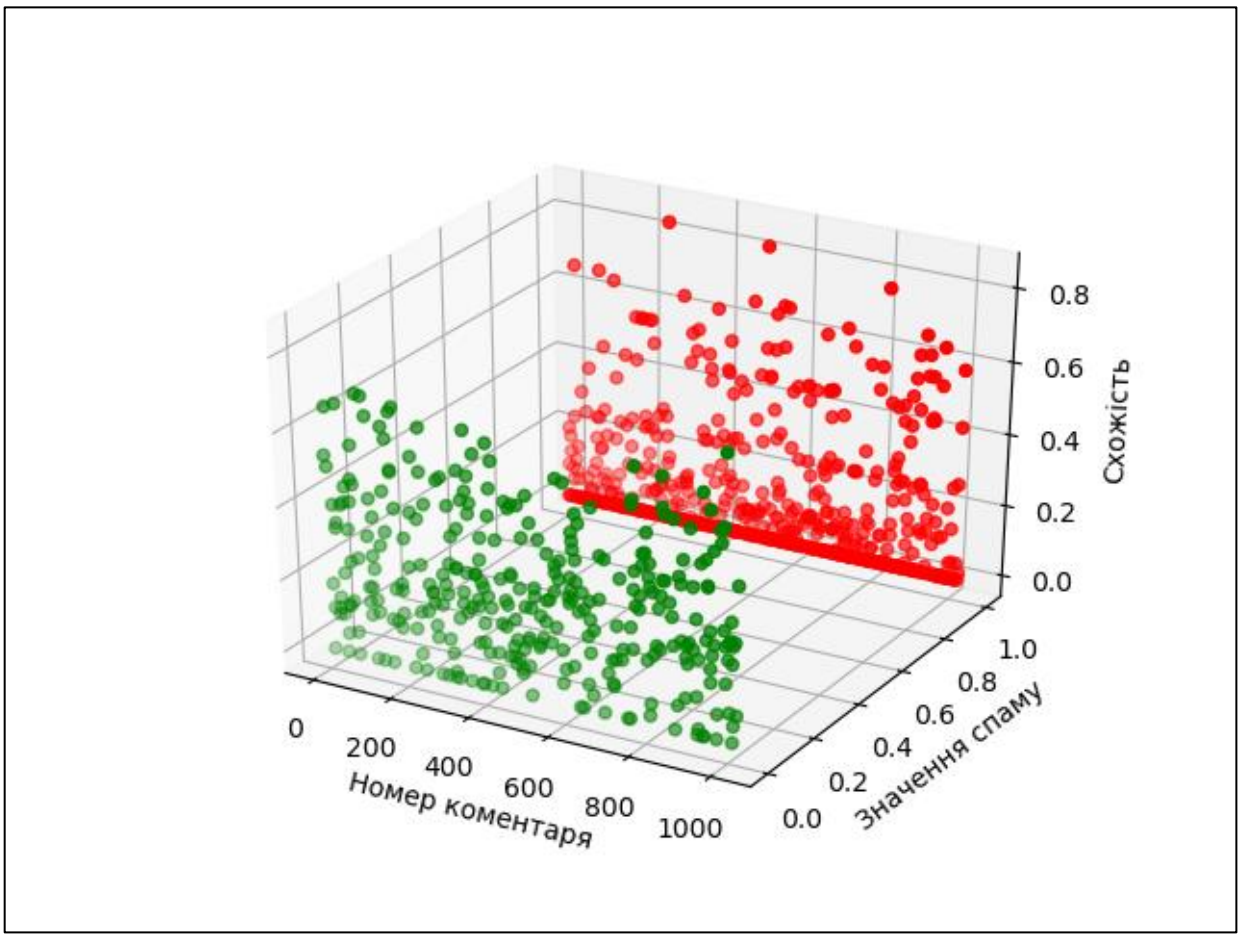

Рисунок 2.1 – Схожість коментаря та статті

За рисунком 2.1, ми можемо точно сказати, що схожість спам-коментарів зазвичай нижча, вона може навіть досягати нульового значення. Цілком покладатися на цей критерій не можна, проте у зв’язках з іншими критеріями він може дати прийнятний результат [7].

2.2 Кількість пробілів

Спам-коментарі зазвичай мають велику кількість пробілів, що йдуть один за одним, для того щоб легше попадатися на очі користувачу. На рисунку 2.2 зображено графік що показує кількість послідовних пробілів для легітимних коментарів та спаму.

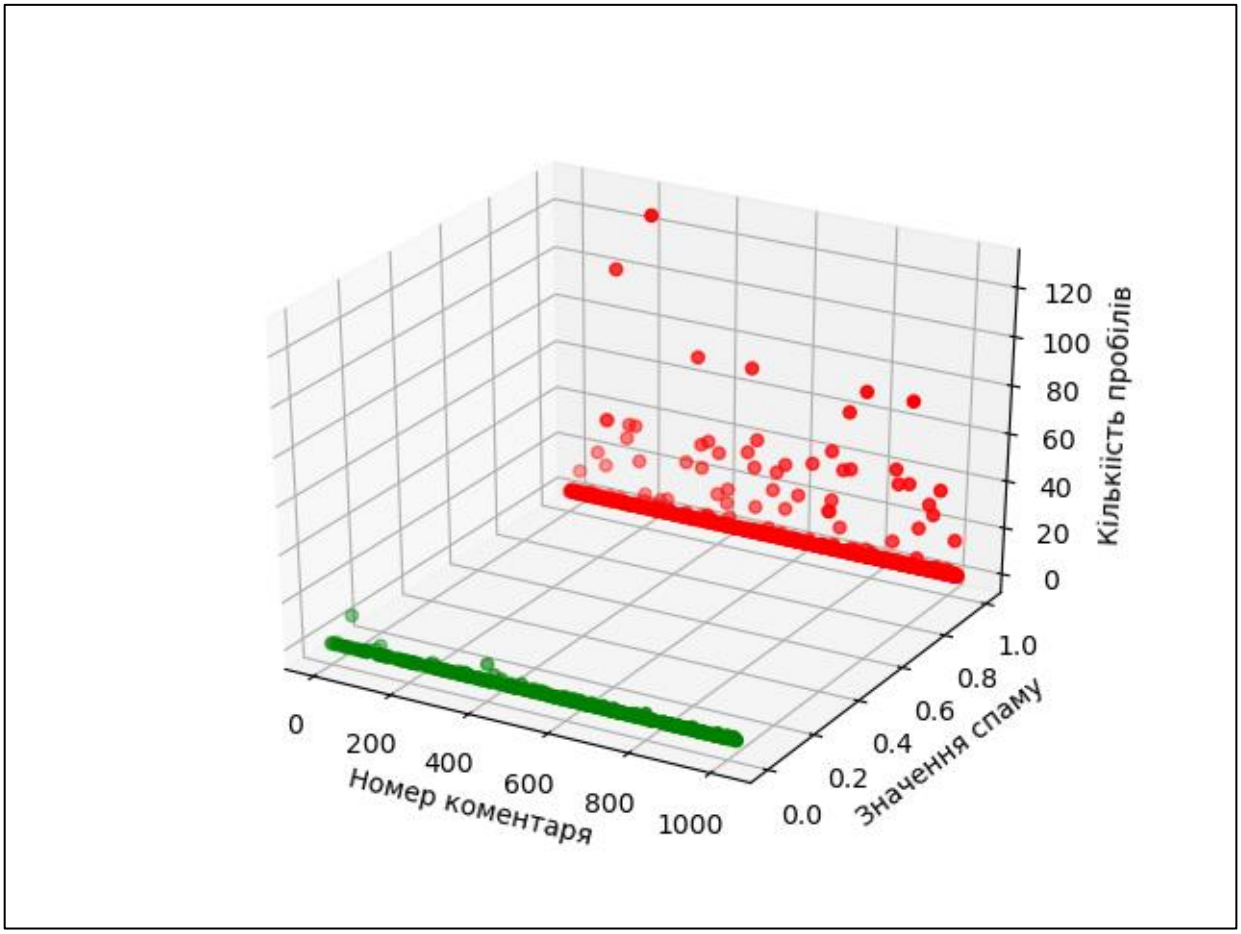

Рисунок 2.2 – Кількість пробілів

З рисунку 2.2 очевидно, що кількість послідовних пробілів в спам-коментарях зазвичай більша. Проте кількість пробілів не може точно сигналізувати нам про наявність спаму, адже вона може бути зумовлена досить довгими реченнями. Саме тому її потрібно застосовувати з критеріями, що розраховують або кількість слів, або кількість речень в коментарі.

2.3 Кількість речень

Кількість речень в спам-коментарі менша за кількість в справжніх, так як справжні коментарі зазвичай мають більшу кількість слів і речення в них розмежовані за правилами. В спам-коментарях навпаки речення не узгоджуються з правилами письма. На рисунку 2.3 зображено графік, що демонструє кількість речень для легітимних коментарів та спаму.

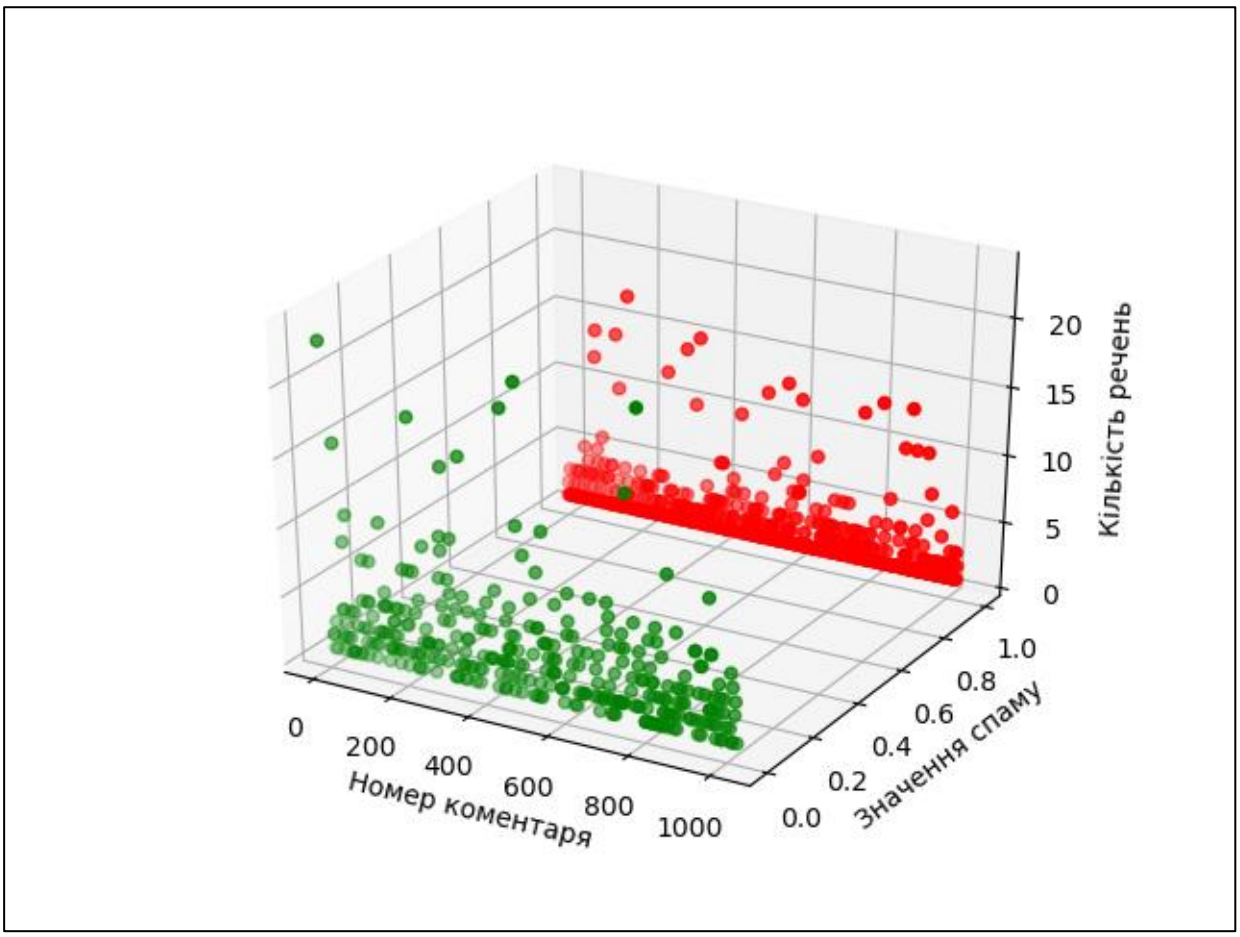

Рисунок 2.3 – Кількість речень

З рисунку 2.3 ми можемо зробити висновок, що наше припущення про більшу кількість речень в коментарі є правдивим, тому ми можемо використати цей критерій в подальших дослідженнях.

2.4 Кількість посилань

Спам-коментарі зазвичай мають більшу кількість посилань, оскільки спамер бажає, щоб користувач перейшов до сайту з малою кількістю відвідувачів. Більш того спамери можуть таким чином рекламувати непопулярні на ринку продукти [4]. На рисунку 2.4 зображено графік, що демонструє кількість посилань для легітимних коментарів та спаму.

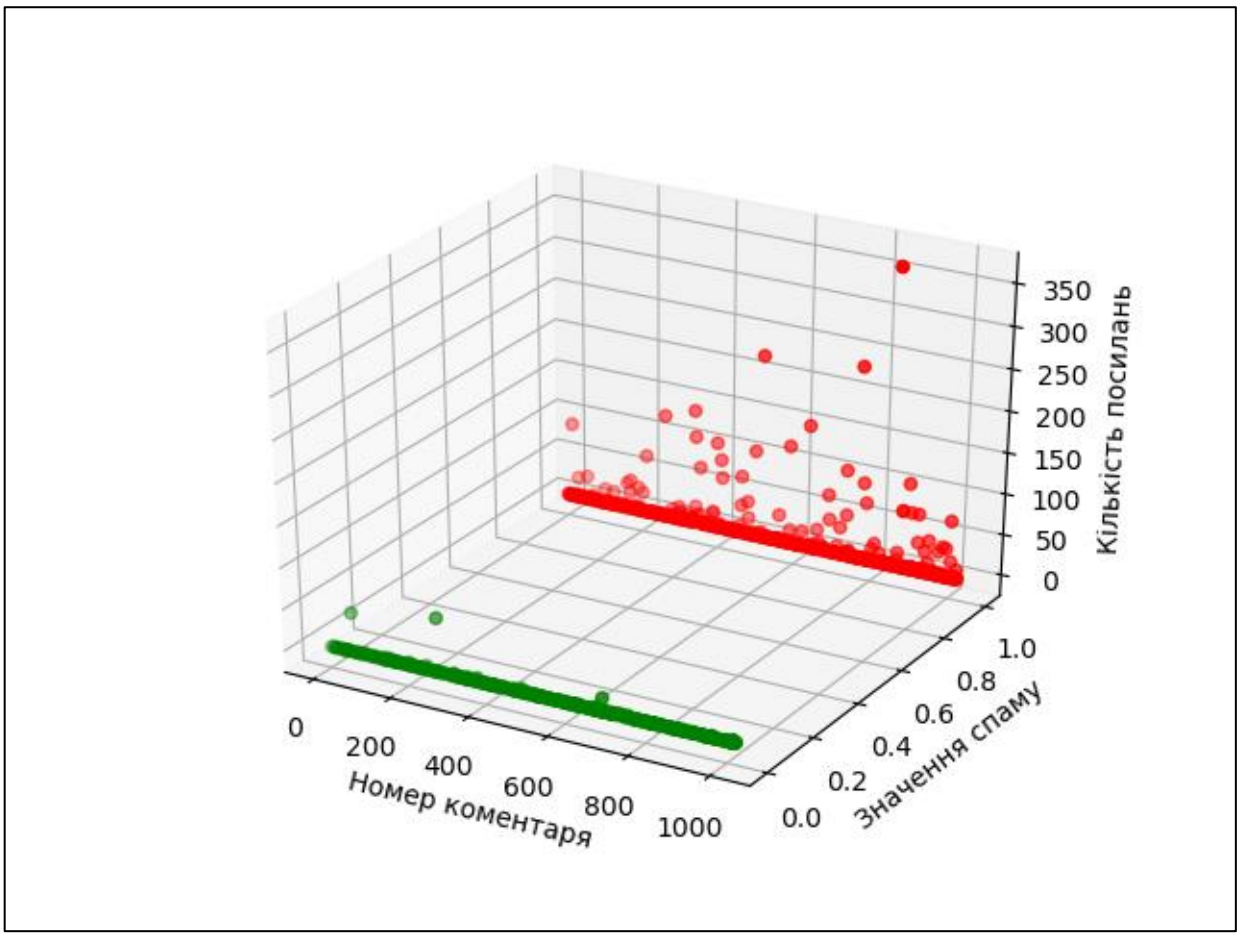

Рисунок 2.4 – Кількість посилань

Кількість посилань – це один з найголовніших критеріїв визначення спаму в інтернеті, адже цей критерій спрямований на пошук найбільш виразної ознаки спаму – посилання на сторонні ресурси. Очевидно, що на рисунку 2.4 ми можемо

бачити колосальну різницю між кількістю посилань в звичайних коментарях та в спамі.

2.5 Пунктуаційні символи

Пунктуаційні символи можуть слугувати досить чітким показником спаму, адже для спам-коментарів дуже характерними є дві ситуації:

– повна відсутність пунктуації;

– наявність підвищеної кількості пунктуаційних символів.

Пунктуаційні символи зазвичай застосовуються спамерами з метою виділення власного коментаря на фоні інших.

В нашому випадку ми вибрали не кількість пунктуаційних символів, а відношення кількості пунктуаційних символів до загальної кількості символів і розраховуємо це значення за формулою 2.2:

$$R_{punctuation} = \frac{N_{punctuation}}{N_{all}} \quad (2.2)$$

де $N_{punctuation}$ – кількість пунктуаційних символів;

N_{all} – кількість всіх символів.

Зазвичай доля пунктуаційних символів відносно всіх інших символів коливається в межах від 0,1 до 0,3. Така стабільність зумовлена особливостями англійської мови і тим, що легітимні коментарі, дуже часто, мають правильну структуру речень, та пунктуацію. І навпаки, коментарі, що підпадають під діапазон від 0 до 0,1 та від0,3 та вище, виглядають підозріло і можуть з високою вірогідністю виявитись спамом.

На рисунку 2.5 зображена доля пунктуаційних символів в коментарях.

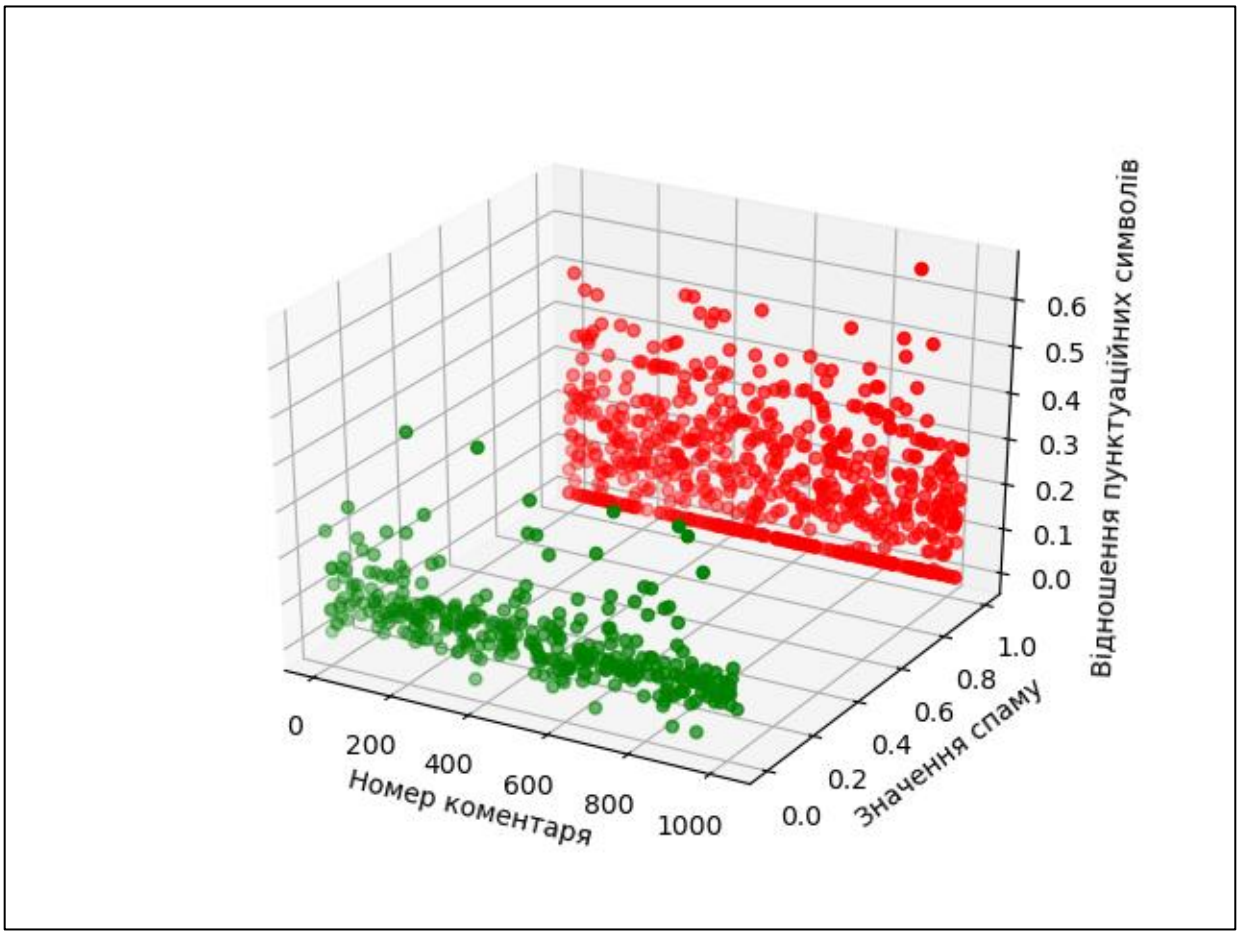

Рисунок 2.5 – Доля пунктуаційних символів

Згідно з рисунком 2.5 ми можемо бачити, що, дійсно, кількість доля пунктуаційних символів в легітимних коментарях коливається в заданих нами межах, в той час як спам-коментарі мають більш широкий діапазон значень. Отже отриманиі дані дають нам підстави використати цей критерій в подальших розрахунках.

2.6 Шумові слова

Шумові слова – це слова, які не несуть смислового навантаження. В англійській мові до таких слів відносять: at, is, the, which та ін. Очевидно, що метою

спам-коментарів є донесення до користувача інформації, отже використання подібних слів є нераціональним.

Доля шумових слів розраховується за формулою 2.3:

$$R_{stop} = \frac{N_{stop}}{N_{all}} \tag{2.3}$$

де N_{stop} – кількість шумових слів

N_{all} – загальна кількість слів.

На рисунку 2.6 продемонстровано відношення кількості шумових слів до загальної кількості слів в коментарі.

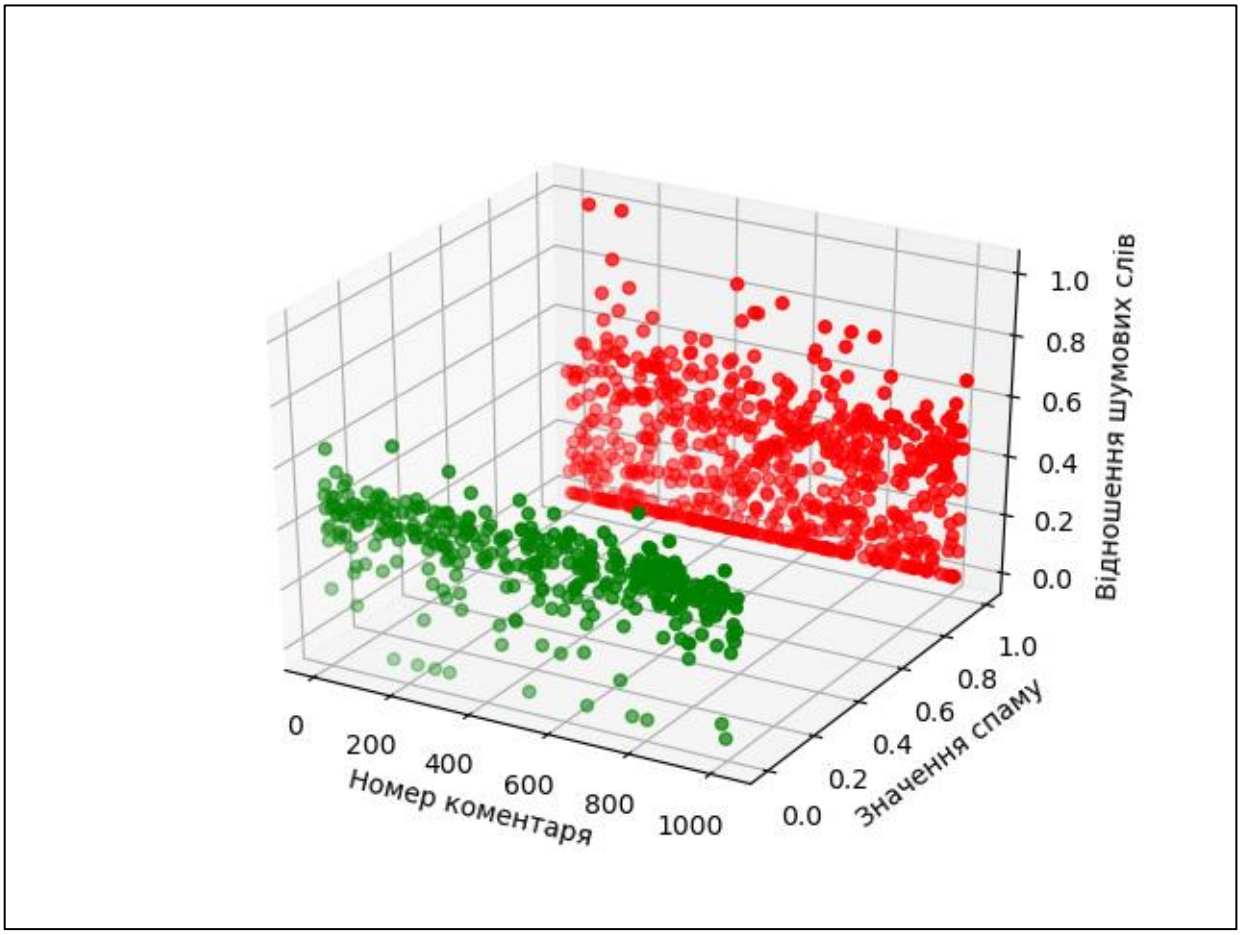

Рисунок 2.6 – Доля шумових слів

Згідно з рисунком 2.6 доля шумових слів більша в легітимних коментарях, в той час як в спамі вона коливається від звичайних значень, притаманних легітимним коментарям і до цілковитої відсутності. Отже цей критерій можна вибрати для пошуку спаму.

2.7 Біграми

Біграма (bigram) – приватний випадок N-gram, що являє собою послідовність двох звуків, складів, слів або букв з точки зору семантичного аналізу. Зазвичай в лінгвістичному аналізі використовують пари: іменник-іменник, прикметник-іменник або іменник-дієслово. Такі пари найбільш точно передають змістовне значення речення.

Для нашого дослідження було прийнято рішення вибрати пари слів, а саме пари іменників, у якості критерія. Це зумовлено тим, що

Згідно з нашим припущенням більшість автоматично згенерованих спам-коментарів насичуються такими парами іменників з метою підвищення ранжування спам сторінок в пошукових системах.

Доля біграм розраховується за формулою 2.4:

$$R_{bigram} = \frac{N_{bigram}}{N_{all}} \tag{2.4}$$

де N_{bigram} – кількість біграм, що складаються з іменників;

N_{all} – загальна кількість слів у коментарі.

На рисунку 2.7 зображено відношення кількості біграм до загальної кількості слів у коментарі.

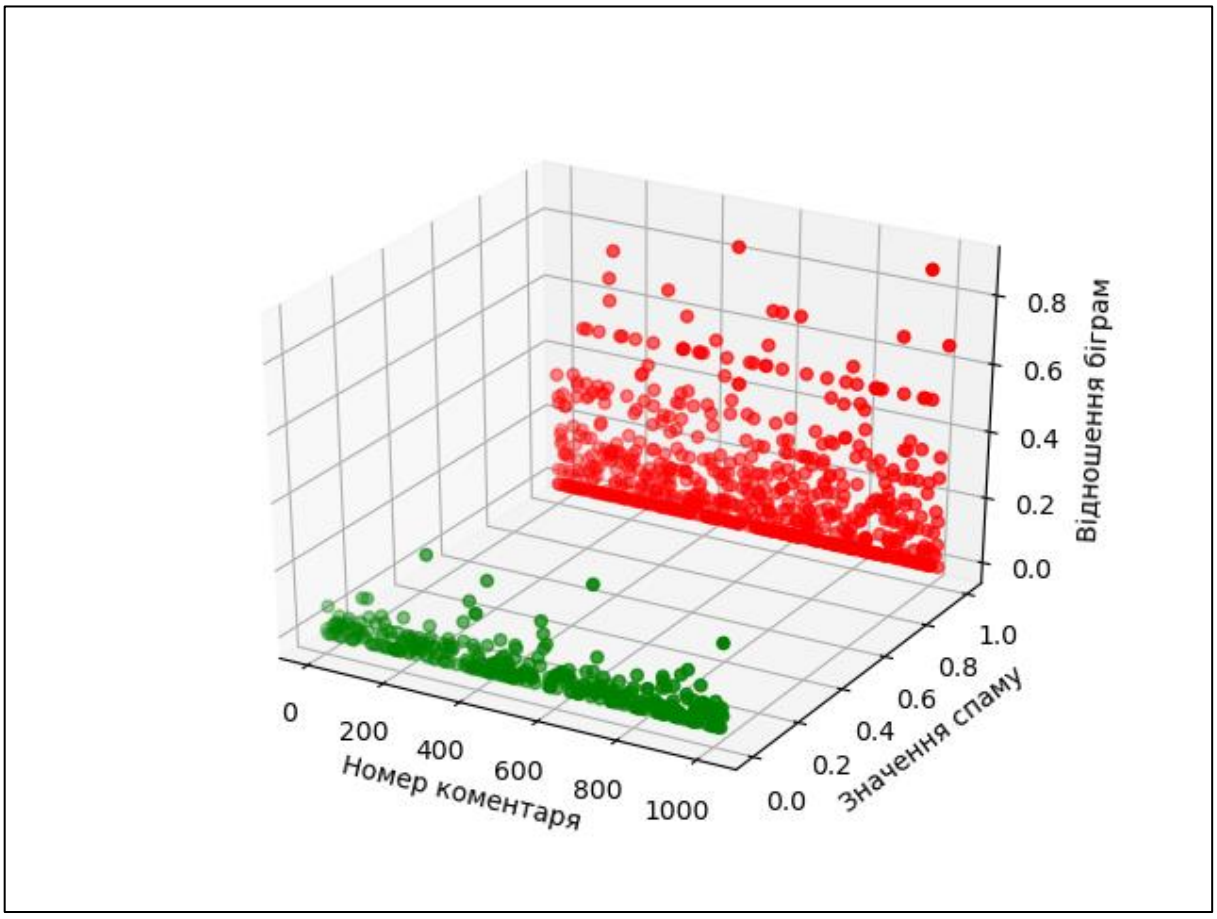

Рисунок 2.7 – Доля біграм

Вочевидь, з огляду на рисунок 2.7 доля біграм, що складаються з іменників, в спам-коментарях є значно більшою. Отже цей критерій можна використовувати в подальших дослідженнях.

2.8 Унікальні слова

Унікальність лексичних засобів – це, зазвичай, те, що відрізняє звичайну людину від машини, чия задача лише формувати коментарі за шаблоном. Спам-коментарі зазвичай намагаються використовувати одні й ті самі слова в порівнянні зі звичайними коментарями, які мають безперервний потік зв`язаного тексту. Доля унікальних слів визначається за формулою 2.5:

$$R_{unique} = \frac{N_{unique}}{N_{all}} \quad (2.5)$$

де N_{unique} – кількість унікальних слів у коментарі;

N_{all} – загальна кількість слів у коментарі.

На рисунку 2.8 продемонстрована доля унікальних слів для легітимних коментарів та спаму.

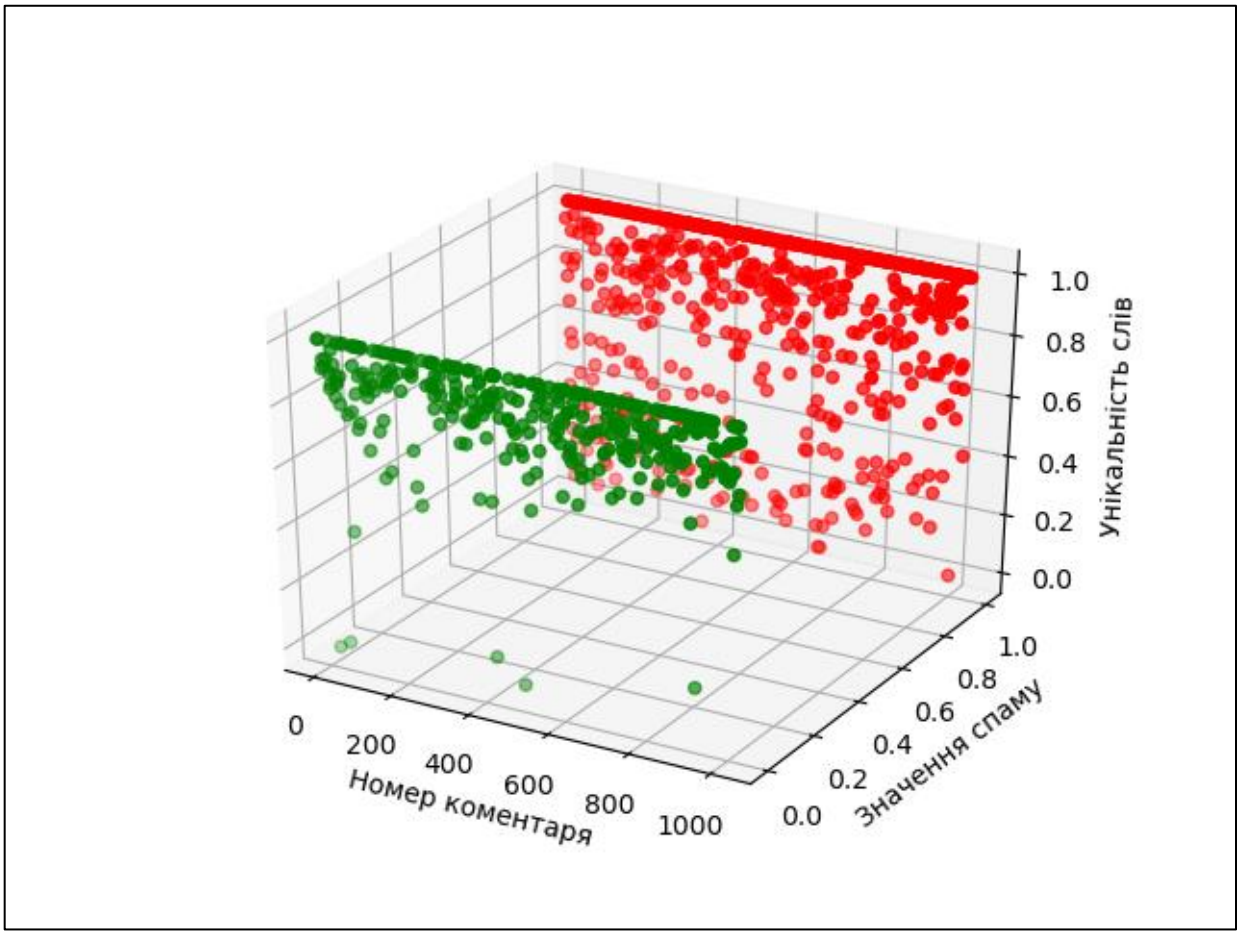

Рисунок 2.8 – Доля унікальних слів

З рисунку 2.8 очевидно, що доля унікальних слів вища у легітимних коментарях, у той час як в спамі вона значно нижча. Отже цей критерій можна використати в подальшому дослідженні.

2.9 Нецензурна мова

Нецензурна мова характерна для спам-коментарів, та й самі нецензурні коментарі можна віднести до категорії спаму. Замість того щоб шукати нецензурні слова було прийнято рішення використати бібліотеку profanity-check для мови Python, щоб визначити вірогідність того, що коментар містить нецензурну лексику. Причиною використання саме цієї бібліотеки є її точність та швидкість.

Точність бібліотеки profanity-check у порівнянні з іншими бібліотеками наведена у таблиці 2.1.

Таблиця 2.1 – Порівняння точності пошуку нецензурної лексики

Бібліотека	Accuracy	Balanced accuracy	Precision	Recall	F1 Score
profanity-check	95,0%	93,0%	86,1%	89,6%	0,88
profanity-filter	91,8%	83,6%	85,4%	70,2%	0,77
Profanity	85,6%	65,1%	91,7%	30,8%	0,46

Швидкість роботи бібліотеки profanity-check у порівнянні з іншими бібліотеками наведена у таблиці 2.2.

Таблиця 2.2 – Порівняння швидкості пошуку нецензурної лексики

Бібліотека	Кількість передбачень		
	1 передбачення	10 передбачень	100 передбачень
profanity-check	0,2	0,5	3,5
profanity-filter	60	1200	13000
Profanity	0,3	1,2	24

На рисунку 2.9 продемонстровано вірогідність наявності нецензурної лексики в коментарі.

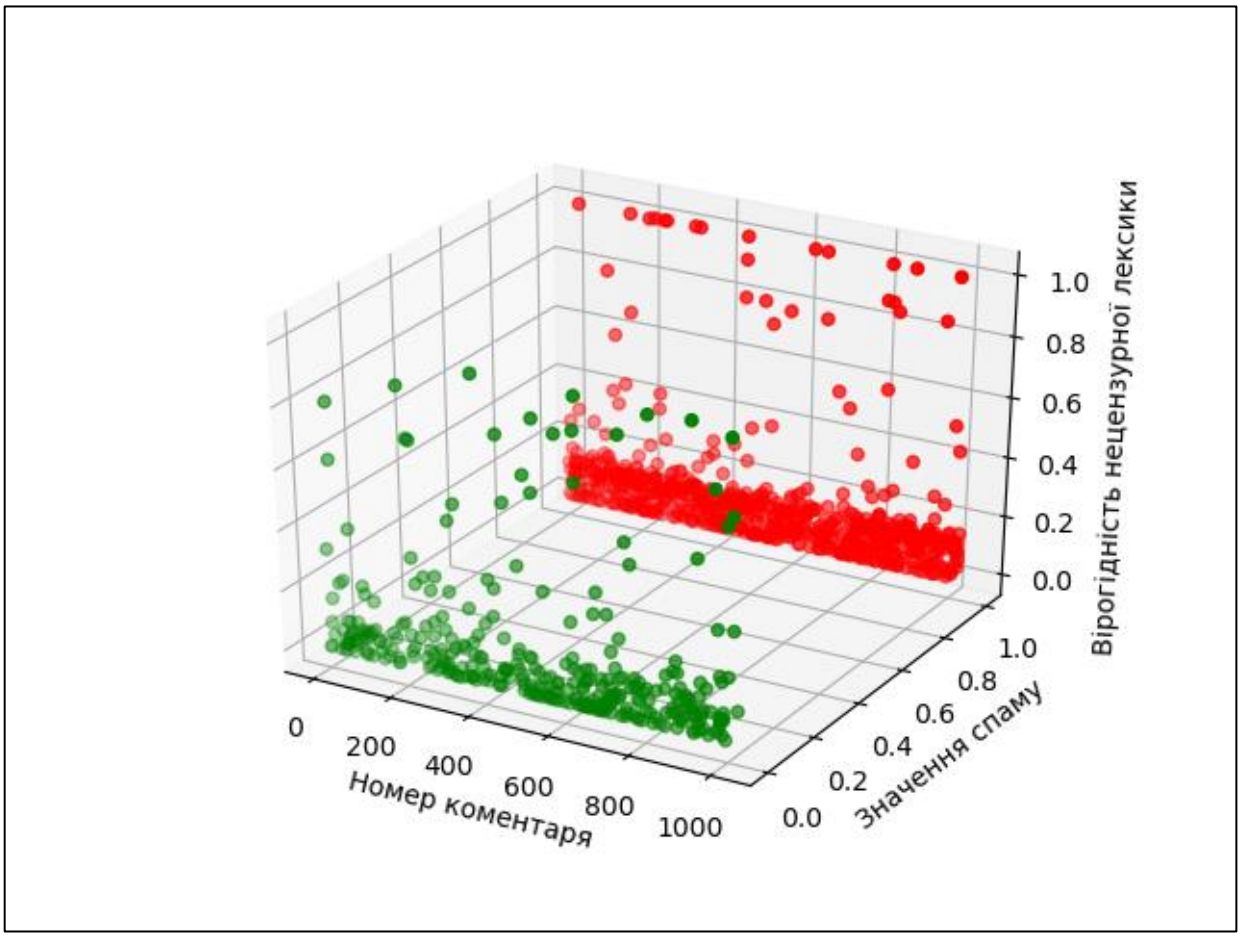

Рисунок 2.9 – Вірогідність нецензурної лексики

Згідно з рисунком 2.9, ми можемо бачити, що спам-коментарі більш схильні мати значення вірогідності нецензурної мови близькі до 1 (100%).

2.10 Загальна характеристика легітимних коментарів та спаму.

Задля того, щоб продемонструвати відмінності в усіх вищеописаних критеріях, ми розробили таблицю для порівняння критеріїв легітимних коментарів та спаму за такими показниками як мода, середнє значення, стандартне відхилення [23]. Результати представлені в таблиці 2.3:

Таблиця 2.3 – Порівняння показників легітимних та спам-коментарів

		Показники					
		Мода		Середнє значення		Стандартне відхилення	
		Спам	Легітимні	Спам	Легітимні	Спам	Легітимні
Критерії	Схожість коментаря та статті	0.0	0.0	0.041	0.275	0.076	0.177
	Кількість пробілів	1	1	3,64	1,18	0,95	10,5
	Кількість речень	1	2	1,95	4,06	2,08	3,61
	Кількість посилань	0	0	6,14	0,6	25,2	4,51
	Доля пунктуаційних символів	0,0	0,1	0,113	0,123	0,118	0,066
	Доля шумових слів	0.0	0.5	0.234	0.427	0.215	0.118
	Доля біграм	0,0	0,0	0,138	0,036	0,165	0,05
	Доля унікальних слів	1.0	1.0	0,808	0,9	0,267	0,131
	Вірогідність нецензурної мови	0,121	0,01	0,114	0,134	0,173	0,211

Згідно з таблицею 2.3 ми можемо простежити відмінності в показниках для легітимних коментарів та спаму для кожного критерія. Надалі необхідно встановити, чи є зв’язок між цими критеріями. Це робиться для того щоб виключити можливість існування одного й того ж критерію під різними назвами. Для цього необхідно встановити кореляцію кожного критерію з кожним, і виходячи з цього виключити або об’єднати сильно залежні між собою критерії.

2.11 Аналіз зв’язку між критеріями

Для того щоб перевірити те, що жоден з вибраних нами критеріїв не дублює інший, ми вирішили застосувати лінійний коефіцієнт кореляції Пірсона. Лінійний коефіцієнт кореляції Пірсона застосовується для аналізу ступеня лінійної залежності між двома змінними.

Результат нашого аналізу записано в таблиці А.1.

Згідно з даними таблиці А.1 ми можемо зробити висновок, що жоден з критеріїв не є лінійно залежним від іншого. Тобто їх всі можна застосовувати для навчання спам-аналізатору. Найбільші рівні залежності проявляються між такими критеріями як: доля шумових слів та доля унікальних слів (0,61) та кількість послідовних пробілів та кількість посилань (-0,58), проте жодне з цих значень не є достатньо високим, щоб стверджувати, що між цими критеріями є лінійний зв’язок.

2.12 Критерії результатів роботи спам-аналізатора

Для того, щоб оцінити результат роботи спам-аналізатора, що використовує для аналізу критерії спаму, що були описані в підпунктах 2.1-2.9 необхідно встановити критерії, за якими буде проводитись оцінка. Для цього були вибрані наступні критерії:

– точність (accuracy, ACC);

– збалансована точність (balanced accuracy, BACC);

– F-score (F1);

– positive predicted value (PPV);

– negative predicted value (NPV).

Для кожного з цих критеріїв встановимо формулу.

Точність дозволяє нам дізнатись відсоток правильно класифікованих коментарів. Точність визначається за формулою 2.6:

$$ACC = \frac{TP+TN}{P+N} \tag{2.6}$$

де TP – кількість коректно класифікованих легітимних коментарів;

TN – кількість коректно класифікованих спам-коментарів;

P – загальна кількість легітимних коментарів;

N – загальна кількість негативних коментарів.

Збалансована точність являє собою середнє значення positive predicted value та negative predicted value, та дозволяє нам дізнатись середнє значення відсотку правильно класифікованих спам-коментарів та правильно класифікованих легітимних коментарів. Збалансована точність визначається за формулою 2.7:

$$BACC = \frac{\frac{TP}{P}+\frac{TN}{N}}{2} \tag{2.7}$$

де TP – кількість коректно класифікованих легітимних коментарів;

TN – кількість коректно класифікованих спам-коментарів;

P – загальна кількість легітимних коментарів;

N – загальна кількість негативних коментарів.

F-score визначається за формулою 2.8:

$$F_1 = \frac{2TP}{2TP+FP+FN} \tag{2.8}$$

де TP – кількість коректно класифікованих легітимних коментарів;

FP – кількість спам-коментарів, класифікованих як легітимні коментарі;

FN – кількість легітимних коментарів, класифікованих, як спам-коментарі.

Positive predicted value визначає відсоток правильно класифікованих легітимних коментарів до загальної кількості легітимних коментарів. Positive predicted value визначається за формулою 2.9:

$$PPV = \frac{TP}{TP+FP} \tag{2.9}$$

де TP – кількість коректно класифікованих легітимних коментарів;

FP – кількість спам-коментарів, класифікованих як легітимні коментарі.

Negative predicted value визначає відсоток правильно класифікованих спам-коментарів до загальної кількості спам-коментарів. Negative predicted value визначається за формулою 2.10:

$$NPV = \frac{TN}{TN+FN} \tag{2.10}$$

де TN – кількість коректно класифікованих спам-коментарів;

FN – кількість спам-коментарів, класифікованих як легітимні коментарі.

Отримавши всі необхідні критерії для формування спам аналізатора та оцінки ефективності його роботи, ми можемо переходити до архітектури та перевірки результатів.

3 АРХІТЕКТУРА ПРОГРАМНОГО РІШЕННЯ

3.1 Загальний огляд архітектурних рішень

Архітектурно система складається з трьох рівні: рівня виділення властивостей, рівня встановлення схожості коментаря та допису та рівня підрахунку результатів.

На рисунку 3.1 продемонстрована загальна архітектура програмного рішення.

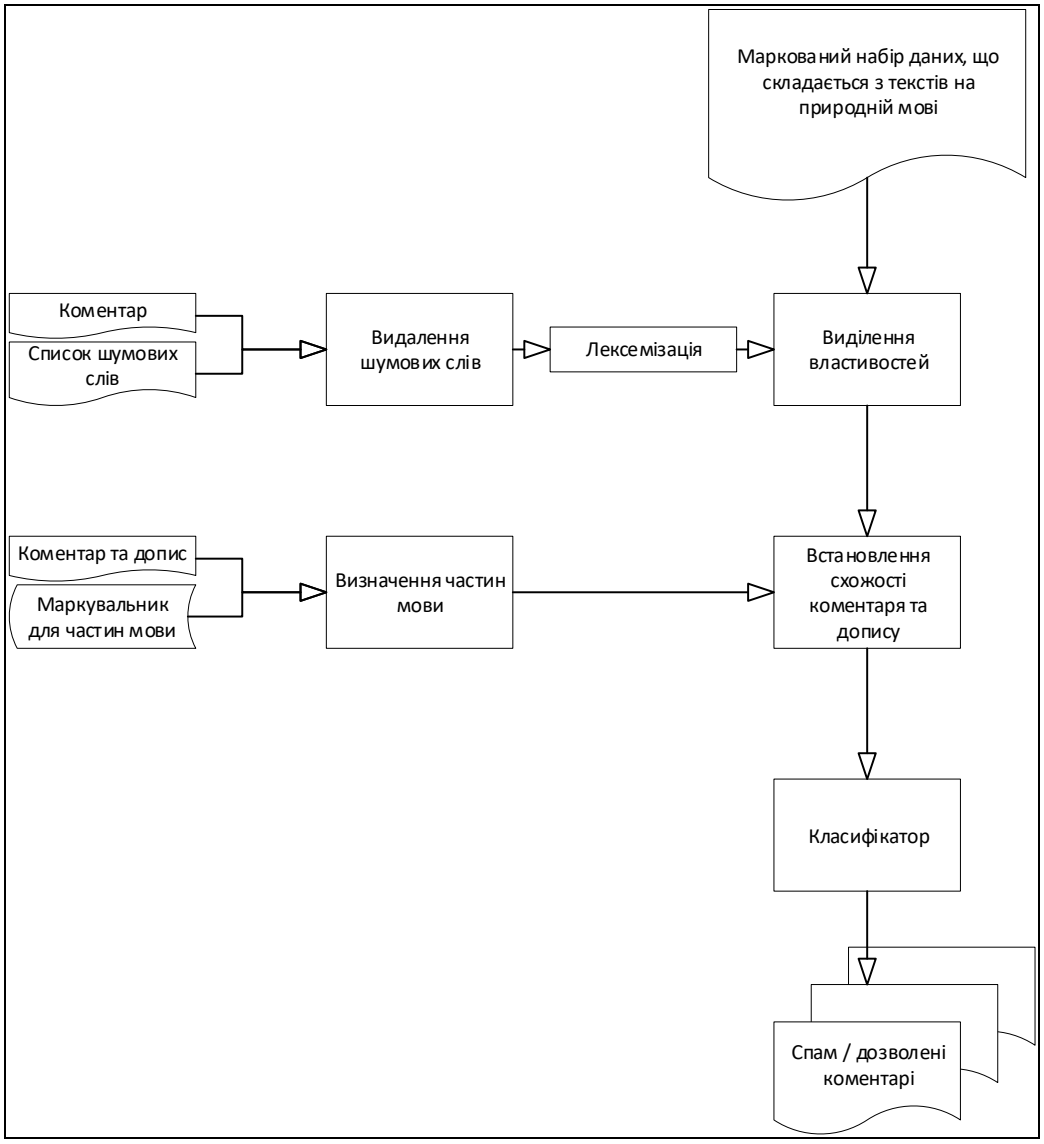

Рисунок 3.1 – Архітектура програмного рішення

На першому рівні нашого рішення відбуваються підготовка коментарів для подальшого аналізу, саме на цьому рівні виділяємо основні кількісні критерії.

На другому рівні ми визначаємо схожість коментаря та допису.

На третьому рівні ми навчаємо нашу систему на основі результатів двох попередніх рівнів та проводимо розрахунки результатів.

Розглянемо рівні більш детально.

3.2 Рівень виділення властивостей.

Так як більшість з наших критеріїв є простими і підходить майже до будь-якої частини тексту, будь то коментар, чи абзац книги, то цей рівень є найбільш насиченим з точки зору визначення властивостей коментаря, що був відправлений на аналіз. Взагалі кажучи, на цьому рівні встановлюються 7 з 9 критеріїв визначених нами раніше.

Спочатку ми отримуємо маркований набір даних, що складається з текстів на природній мові, та список шумових слів. На цьому етапі ми приводимо всі слова в текстах до єдиного вигляду (приводимо до нижнього регістру). Після чого відбувається підрахунок проміжних результатів, серед них: підрахунок кількості url посилань та речень. Відтак, ми видаляємо всі посилання та теги з коментаря. На наступному етапі ми підраховуємо кількість послідовних пробілів (два або більше, що йдуть один за одним). Також ми визначаємо вірогідність того, що в тексті була використана нецензурна мова. Після чого відбувається лексичний аналіз і текст розбивається на окремі слова. Згодом, відбувається підрахунок долі пунктуаційних символів та їх видалення. Таким же чином розбираємось і з шумовими словами. І наостанок підраховуємо критерій унікальності слів.

На рисунку 3.2 продемонстрована схема першого рівня архітектури.

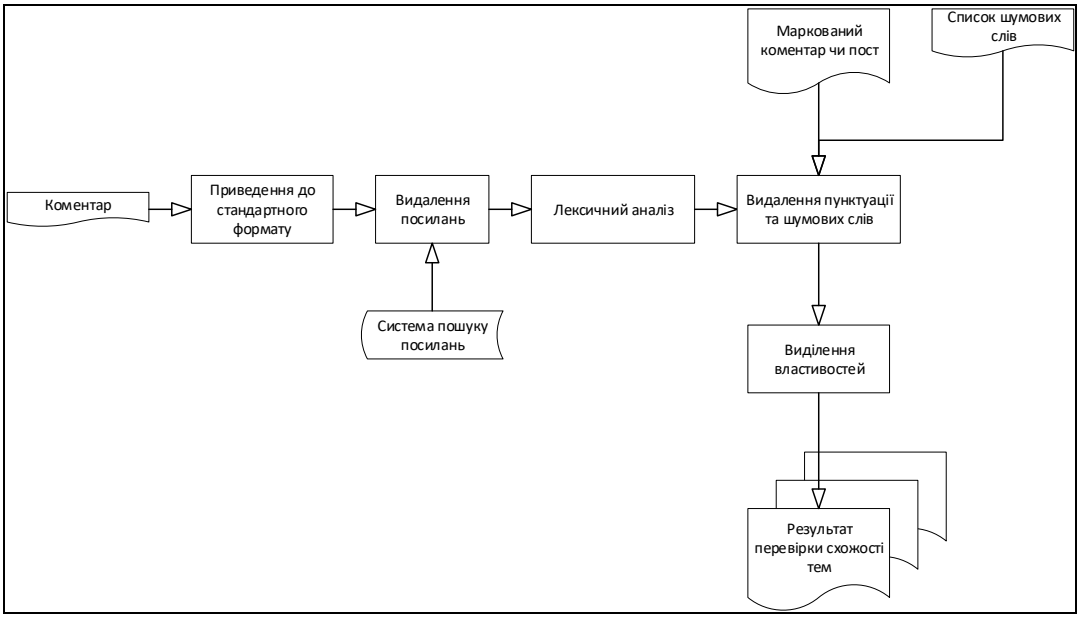

Рисунок 3.2 – Схема роботи рівня виділення властивостей

Після проходження рівня виділення властивостей ми отримуємо коментар, що готовий для подальшої обробки та використання на рівні встановлення схожості коментаря та допису.

3.3 Рівень визначення схожості коментаря та допису

На цьому рівні, ми, використовуючи результати попереднього етапу визначимо останні два критерія.

Послідовність роботи цього рівня наступна: спочатку ми визначаємо частину мови для кожного елементу в нашому обробленому коментарі. Після чого ми розраховуємо значення долі біграм, що складаються з іменників, в нашому коментарі. І останнім ми визначаємо схожість коментаря та допису. В результаті ми отримуємо числову характеристику, за допомогою якої можемо встановити схожість коментаря та допису.

На рисунку 3.3 зображено схему роботи рівня визначення схожості коментаря та допису.

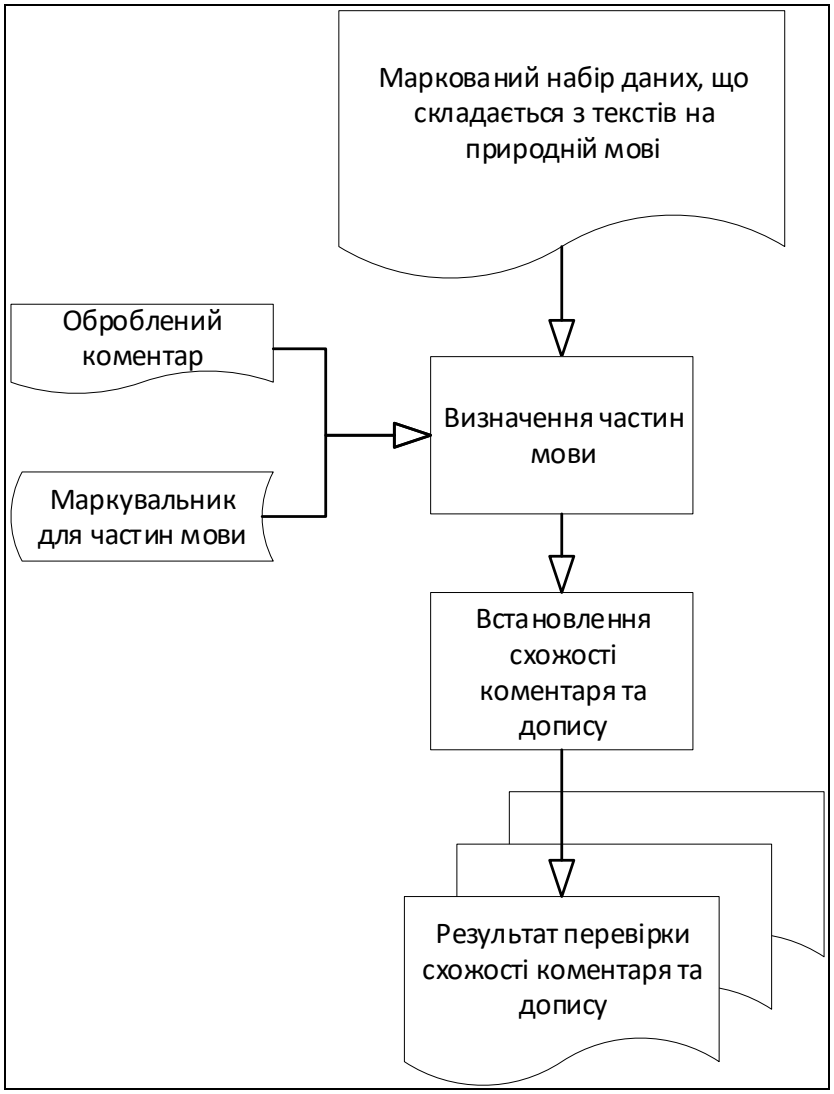

Рисунок 3.3 – Схема роботи

В результаті, після проходження цього етапу ми маємо всі необхідні дані для класифікації нашого коментаря і можемо переходити до етапу класифікації.

3.4 Етап класифікації

На етапі класифікації, ми, використовуючи класифікатор на основі методу опорних векторів встановлюємо чи є розглянутий на попередніх етапах коментар спамом, чи він цілком легітимний. Класифікатор попередньо навчається на

заздалегідь підготованому набору даних. Якщо коментар виявляється спамом, то результатом буде 1, якщо він легітимний, то результат 0.

3.5 Результати досліджень

Дослідження було проведено на наборі даних, що складається з 1024 коментарів, з яких 332 легальні, а інші 692 – спам.

Після навчання спам-аналізатору на основі методу опорних векторів та критеріїв спаму, що були описані в підпунктах 2.1-2.9, ми провели його перевірку. За результатами цієї перевірки виявилось, що:

– правильно визначених, як спам, коментарів (True Negative, TN) 660;

– неправильно визначених як спам, коментарів (False Negative, FN) 101;

– правильно визначених, як легітимні, коментарів (True Positive, TP) 231;

– неправильно визначених, як легітимні, коментарів (False Positive, FP) 32.

Тоді згідно з критеріями визначеними в підпункті 2.12 сформуємо таблицю 3.1.

Таблиця 3.1 – Оцінка роботи класифікатора

	Критерії				
	NPV	PPV	F_1	ACC	BACC
Результати	86,7%	87,8%	90,8%	87%	82%

В результаті дослідження було сформовано цілком працездатну систему, що може класифікувати коментарі з точністю 87%, що є досить високим показником, приймаючи до уваги те, що в роботах попередників, найкращим з досягнутих результатів була точність в 92%.

ВИСНОВКИ

В ході атестаційної роботи було проведено дослідження методів лінгвістичного аналізу відгуків користувачів інтернет-форумів. В ході самого дослідження, було доведено актуальність розглянутої проблематики. Також, спираючись на методи лінгвістичного аналізу було сформовано унікальні критерії спаму. Основною особливістю більшості з цих критеріїв є їх універсальність, що дає змогу застосувати їх до будь-якого коментаря.

Спираючись на роботи попередників та отримані критерії спаму була сформована трьохрі0432нева архітектура спам-аналізатора. Після чого було розроблено прототип системи для перевірки його працездатності та точності.

Перевірка проводилась на датасеті, створеному Мішне та Кармел, що складається з 50 дописів та 1024 коментарів до цих дописів. За її результатами виявилось, що точність створеного нами аналізатору на основі методу опорних векторів становить 87%. У порівнянні з результатами попередніх робіт, що були розглянуті, він займає друге місце за точністю та поступається лише роботі Хуана та ін. з їх 92% точності. Було встановлено також, що система краще ідентифікує легітимні коментарі (з точністю 87,8%) на 1,1% порівняно зі спам-коментарями (86,7%).

Результати роботи можна застосувати в подальшому вивченні проблем спаму та для пошуку інших універсальних критеріїв, що характеризують спам.

У планах перевірка продуктивності роботи розробленої системи з використанням більших за розміром об’ємів даних. Також, переконаний, що описаний в роботі підхід можна застосувати для роботи зі спамом в YouTube, Twitter та ін. системах.

ПЕРЕЛІК ПОСИЛАНЬ

1. M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, “A Bayesian approach to filtering junk e-mail”. In AAAI-98 Workshop on Learning for Text Categorization, July 1998, Madison, Wisconsin, pp. 98-105.

2. Gilad Mishne, David Carmel, and Ronny Lempel, “Blocking blog spam with language model disagreement”. In Proceedings of the First International Workshop on Adversarial Information Web (AIRWeb), Chiba, Japan, May 2005, pp. 1-6.

3. A. Bhattari and D. Dasgupta, “A Self-supervised Approach to Comment Spam Detection based on Content Analysis”. In International Journal of Information Security and Privacy (IJISP), Volume 5, Issue 1, 2011, pp. 14-32.

4. Davison B. D., “Recognizing Nepotistic Links on the Web”. In AAAI 2000 Workshop on Artificial Intelligence for Web Search, 2000, pp. 23- 28

5. Carreras, X. and Marquez, L., “Boosting trees for anti-spam email filtering”. In Proceedings of RANLP-01, 4th International Conference on Recent Advances in Natural Language Processing, 2001, pp. 58-64.

6. Stefan Siersdorfer and Sergiu Chelaru, “How useful are your comments?: analyzing and predicting YouTube comments and comment ratings”. In Proceedings of the 19th international conference of World wide web, 2010, pp. 891-900.

7. Ruihai Dong, Markus Schaal and Barry Smyth, “Topic extraction from online reviews for classification and recommendation”. In Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, 2013, pp. 1310-1317.

8. The Stanford Natural Language Processing Group URL: http://nlp.stanford.edu/software/corenlp.shtml (дата звернення: 20.10.19).

9. Serbanoiu, A., Rebedea T., “Relevance-Based Ranking of Video Comments on YouTube”. In CSCS ’13 Proceedings of the 2013 19th International Conference on Control Systems and Computer Science, 2013, Washington, USA, pp. 225-231.

10. I. Drost and T. Scheffer., “Thwarting the nigritude ultramarine: Learning to identify link spam”. In ECML’05 Proceedings of the 16th European conference on Machine Learning, 2005, Berlin, Germany, pp. 96-107.

11. Zhang, L., Zhu, J., & Yao, T. (2004). An evaluation of statistical spam filtering techniques. ACM Transactions onAsian Language InformationProcessing, 3(4), 243–269. doi:10.1145/1039621.1039625.

12. Akismet. (n.d.). URL: http://akismet.com/ (дата звернення: 20.10.19)

13. Drucker, H., Wu, D., & Vapnik, V. (1999). Support vector machines for spam categorization. IEEE TransactionsonNeuralNetworks, 10(5), 1048–1054. doi:10.1109/72.788645.

14. Becchetti, L., Castillo, C., Donato, D., Leonardi, S., & Baeza-Yates, R. (2005). Link-based Characterization and Detection of Web Spam. In Proceedings of the 2nd International Workshopon Adversarial Information Retrieval on the Web (AIRWeb), Seattle, WA.

15. Umbria. (2005). Spamin the blogosphere. URL: http://uploadi.www.ris.org/editor/1135776405umbria_splog.pdf (дата звернення: 20.10.19).

16. Gyongyi, Z., & Hector, G. (2005). Web Spam Taxonomy. In Proceedings of the Adversarial Information Retrieval on the web (AIRWeb) Conference (pp. 39-47).

17. Ntoulas, A., Najork, M., Manassee, M., & Fetterly, D. (2006). Detecting Spam Web Pages through Content Analysis. In Proceedings of the WWW 2006 Conference, Edinburgh, UK.

18. Kolari, P., Java, A., Finin, T., Oates, T., & Joshi, A. (2006). Detecting Spam Blogs: A Machine Learning Approach. In Proceedings of the AAAI 2006 Conference.

19. Han, S., Ahn, Y. Y., Moon, S., & Jeong, H. (2006). Collaborative Blog Spam Filtering Using Adaptive Percolation Search. In Proceedings of the WWW2006 Conference, Edinburgh, UK.

20. Cormack, G. V., Gomez, J. M., & Sanz, E. P. (2007). Spam Filtering for short messages. In Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM 2007).

21. YouTube Community Guidelines enforcement URL: https://transparencyreport.google.com/youtube-policy/removals?comments_by_source=period:Y2019Q3&lu=total_channels_removed&total_comments_removed=period:Y2018Q3&channels_by_reason=period:Y2019Q3&videos_by_reason=period:Y2019Q3&content_by_flag=period:Y2019Q3;exclude_automated:&total_channels_removed=period:Y2019Q3&hl=en (дата звернення: 20.10.19).

22. Spam and phishing in Q3 2019 URL: https://securelist.com/spam-report-q3-2019/95177/ (дата звернення: 20.10.19).

23. Н.В. Голян, В.В. Голян. Modern technologies for collectiom and processing of a great amount of big data//Monograph. Big Data Processing: Metothods, Models and Information Technologies. Shioda GmbH. Steyr, Austria, 2019 – 26 -34.