ШИФР **<u>ІФБР-4</u>**

ЗАСТОСУВАННЯ МОДЕЛІ МНОЖИННОЇ ЛІНІЙНОЇ РЕГРЕСІЇ У ПРОГНОЗУВАННІ ПОПУЛЯРНОСТІ УКРАЇНСЬКОМОВНОГО INSTAGRAM-ТЕКСТУ

ЗМІСТ

ВСТУП

Проникнення Інтернету в усі сфери діяльності людини, зокрема й у повсякденну комунікацію, визначило пріоритетними об’єктами лінгвістичного дослідження комп'ютерно-опосередковану мовну діяльність і текст в електронній формі – результат цієї діяльності. Такий підхід спричинив експансіонізм нової лінгвістичної парадигми й інтеграцію з іншими науками, у зв’язку з чим виникло багато нових предметів і методів міждисциплінарних досліджень, які належать до сфери прикладної лінгвістики. Особливе місце у поліпарадигмальному дискурсі сучасної лінгвістики належить міжгалузевому синтезу інтернет-лінгвістики та комп’ютерної лінгвістики, яка виконує функцію процесуально-технічного складника у всіх галузях сучасного лінгвістичного пізнання, керованого прикладними завданнями.

Нова форма спілкування – інтернет-комунікація – стала об’єктом вивчення великої плеяди мовознавців таких, як М. Бергельсон [2], Т. Іванова [5], Л. Компанцева [7], В. Фатурова [16], Г. Чичкань [14], С. Зайцева [51], Ю. Щуріна [40] та інших, які визначають інтернет-комунікацію новою формою міжособистісного спілкування, що характеризується глобальністю масштабів і можливістю миттєвого вільного поширення будь-якої інформації, а також зміною прагматичних установок та цілей учасників цього виду комунікації, актуалізацією феномену мовленнєвої особистості, яка через посередництво системи мовленнєвого інтернет-жанру реалізує себе у віртуальному дискурсі.

Однією з найпопулярніших інтернет-платформ в інформаційно-комунікативній віртуальній реальності О. Гримов [25] називає соціальні мережі, які, на думку Ж. Денисюк [21], не тільки задовольняють комунікативні потреби людини, а й розширюють свої функції, надаючи користувачеві адаптивні життєві стратегії, що насамперед зумовлені прагненням до популярності. Тому сучасна інтернет-комунікація, що використовує маркетингові прийоми популяризації користувача в мережі М. Окландер [8],

стала монетизованим фактором, який має великий вплив на комунікативну та споживчу поведінку користувача. Індикатором такої поведінки став персональний вплив «лідерів громадської думки» (Н. Савицька [10], Сет Годін [24]), маркетингова стратегія яких ґрунтована на своїй популярності й поширенні свого стилю споживчої поведінки на цільові аудиторії. Такими лідерами в одній із найпопулярніших соціальних мереж Instagram є інстаблогери – люди, що ведуть облікові записи й наростили велику кількість читацької аудиторії за допомогою створеного фото-, відео- та текстового контенту. Інстаблогери надають рекламні послуги в межах власної читацької аудиторії, просуваючи бренди, продукти, товари тощо. Тому сьогодні якомога більше людей намагаються популяризувати свої облікові записи в Instagram й наростити читацьку аудиторію для залучення рекламних пропозицій.

Популярність облікового запису (або блогу) Instagram визначається трьома показниками: кількістю читацької аудиторії облікового запису, кількістю вподобань під публікаціями облікового запису та статистикою відвідувань облікового. Аудиторія інстаблогера може підвищувати популярність його блогу, ставлячи вподобання (лайк) та коментар під публікацією

Спочатку Instagram був лише платформою для публікування фотографій без підпису, тому більшість методик підвищення популярності блогу базується на виборі сенсаційного фото або відео. Щоб зробити соціальну мережу більш цікавою широкому загалу, з часом виникла функція підписування фотографії – створення текстового полікодового контенту (посту), що поєднує вербальні та невербальні засоби інтернет-комунікації. Багатьом інстаблогерам вдалося здобути популярність та монетизувати свій блог за допомогою не тільки цікавої фотографії, а й вдалої комунікативної стратегії в організації текстового Instagram-дискурсу.

Вивчення факторів популярності веб-сайтів за допомогою комбінацій ключових слів і фраз, якісного контенту й пов'язаного з ними рейтингу за

даними пошукових систем належить до завдань відносно молодої міжгалузевої прикладної науки, що називається SEO-копірайтинг: SEO (від англ. *search engine optimization*) – пошукова оптимізація сайту; копірайтинг (від англ. *copywriting*: *copy* – рукопис, текстовий матеріал; *write* – писати). SEO-копірайтинг – професійна діяльність з написання рекламних та презентаційних текстів, а також наука, яка вивчає методи цієї діяльності. SEO-копірайтинг ще називають мистецтвом створювати магніт для трафіку, але сучасні аналітичні системи у сфері SEO спрямовані на аналіз великих оригінальних текстів: оглядів, аналітики, описів товарів і послуг. Короткі тексти соціальних мереж, зокрема instagram-тексти, обмежені обсягом 2200 символів, можуть аналізуватися лише за зовнішніми факторами пошукової оптимізації, тому що статистичні методи визначення ключових слів та метод латентного семантичного аналізу не є ефективними для аналізу коротких текстів. На сьогодні ще не існує систем для комплексного SEO-аналізу, мінімальних за обсягом, текстових публікацій у популярній соціальній мережі Instagram. Вивчення текстових особливостей цього дискурсу, а також текстових і позатекстових факторів, що впливають на популярність instagram-блогу – актуальне завдання сучасної комп'ютерної лінгвістики.

Мета дослідження – створити автоматичну систему визначення факторів впливу на популярність текстової публікації у соціальній мережі Instagram.

Досягнення мети передбачає виконання таких **завдань:**

1) сформувати корпус текстових instagram-публікації;

2) визначити лінгвістичні та структурні параметри, які зумовлюють популярність текстової публікації;

3) розробити програмне забезпечення для автоматичного визначення у текстах лінгвістичних та інформаційних параметрів, які можуть мати вплив на популярність текстової публікації;

4) розробити комп'ютерну програму автоматичного визначення прогностичних параметрів популярності текстової instagram-публікації на основі моделі множинної лінійної регресії.

Об’єктом дослідження є блогова українськомовна instagram-публікація.

Предметом дослідження є інформаційні та лінгвістичні фактори популярності українськомовного instagram-тексту.

Методи дослідження: методи статистичного аналізу, метод машинного навчання, метод моделювання множинної лінійної регресії, методика проведення тонального аналізу тексту.

Інформаційна база дослідження: мова програмування Python та її бібліотеки: emoji [55], lxml [56], matplotlib [57], mpl_toolkits [61], nltk [59], numpy [58], pandas [60], pymorphy2 [53], requests [62], sklearn [54], sqlite3 [64], tokenize_uk [63].

Матеріал дослідження: 500 текстових полілогів облікового запису @nata_fedorchuk із соціальної мережі Instagram.

РОЗДІЛ 1
ЛІНГВІСТИЧНІ ОЗНАКИ ТА ІНФОРМАЦІЙНІ ОСОБЛИВОСТІ ІНТЕРНЕТ-ТЕКСТУ

Інтернет-комунікація зумовила появу нового лінгвістичного явища – інтернет-тексту, що, на думку В. Чернявської [13], характеризується медійністю за способом передачі інформації, зокрема технічно-опосередкованим комунікативним каналом. Поняття медійності, за визначенням Є. Гончарової [4], також пов'язане із поняттям комунікативного коду – системою знаків, символів та правил їхнього поєднання для передачі, оброблення й зберігання інформації в зрозумілій формі. Комунікативний код та комунікативний канал можуть поєднуватися настільки сильно, що деякі елементи комунікативного каналу стають елементами комунікативного коду. Тому техногенні процеси в сучасній комунікації та виникнення нових способів і форм передачі інформації спричинили виникнення «медійного перевороту». Медійність змінює співвідношення змісту тексту та форм його вираження. Специфіка медійної форми впливає на формування, структурування й презентацію змісту, а також на його сприйняття адресатом.

В. Чернявська [12] вводить поняття полікодовості на позначення взаємодії різних комунікативних каналів. Вивчення цього поняття вимагає поєднання міжгалузевих методик аналізу тексту: культурології, семіотики, теорії комунікації тощо. У полікодових текстах поєднані компоненти мовних та немовних кодів, що обумовлено мультимедійністю засобів масової комунікації. Тому сучасні дослідження тексту не можуть не враховувати його медійного аспекту, а поява полікодових текстів є наслідком візуалізації комунікативного повідомлення та намагання надати комунікації більш естетичного вигляду, що, на думку У. Фікс [22], означає акцентуацію тексту на собі через виділення та підкреслення власної форми. «Сьогодні естетизація, тобто опора на зовнішню красу, дизайн, характеризує практично всі форми комунікації: рекламу, ЗМІ, політику, сферу щоденного спілкування… Традиційно до засобів естетизації

належать мовні, а саме стилістичні, риторичні аспекти «красномовства». Крім того, сьогодні очевидно, що значний естетичний потенціал міститься в матеріальній організації текстів. Це означає посилену увагу до їх форми, гри з формою. Форма тексту стає додатковим засобом його виділення в загальному інформаційному просторі. Форма стає тим маркером, який забезпечує максимальну концентрацію уваги на своєму об'єкті» [13, 70 – 71].

Трактування полікодовості тексту не зводиться до поняття його візуалізації, адже техногенні процеси у сфері комунікації призвели до ускладнення самого феномену тексту. У полікодовому тексті візуальні та вербальні знаки комунікації формують смислове ціле. Візуальне та вербальне можуть поєднуватися між собою додатковими та інтегральними відношеннями. У додаткових відношеннях візуальне та вербальне виражають один і той самий зміст, проте візуальне є засобом подвійного кодування вербального змісту. В інтегральних відношеннях неможливо відділити вербальне від візуального, хоча можливе семантичне доповнення, коли візуальний компонент робить свій внесок у створення когерентного цілого разом з іншими елементами текстової структури.

Інтернет-комунікація зумовила створення відкритого дискурсу, а отже й нових жанрів комунікації, які спричинили руйнування меж тексту: текст необхідно розглядати не автономно, а в інтернет-просторі загалом. Одним із таких жанрів інтернет-комунікації є блог. Його особливість – відкритість полілогового дискурсу текстової публікації автора, яка визначає тему полілогу, а коментарі до цієї публікації формують відкритий полілог.

Полілоговий дискурс інтернет-комунікації характеризує ще одна ознака – монетизація. Цей аспект пов’язаний із критерієм популярності тексту, що визначається кількістю учасників полілогу. Це вимагає від автора пристосовуватися до своєї читацької аудиторії й писати тексти на актуальні теми. Тому виникла ідея створення системи, яка б стала для авторів блогів надійним помічником у створенні популярного інтернет-тексту.

Полікодовий текст об’єднує в собі вербальні та невербальні засоби передачі інформації і потребує вивчення обох аспектів текстової структури. Д. Крістал**,** описуючи мову інтернет-спілкування, пропонує таку формулу його структури: «усна форма мови + письмова форма мови + ознаки, опосередковані комп’ютером» [20, 33]. До ознак усного мовлення, що зустрічаються у полікодових текстах, належать емоційність, невимушеність, перерваність, логічна непослідовність висловлювань. Зокрема, прагнення до емоційності висловлювання виражається за допомогою піктограм. За теорією Ч. Пірса [6], піктограми – це знаки-ікони, які представляють зовсім інший спосіб кодування інформації, ніж вербальні знаки-символи. Полікодовий текст, який використовує піктограми, набуває нових семіотичних ознак, тому що поєднує лінгвістичні знаки-символи з іконічними знаками. Особливостями писемного мовлення в інтернет-комунікації є порушення норми письмової літературної мови та нівелювання кодифікованих етикетних формул, що зумовлено насамперед анонімністю спілкування та швидкістю комунікативного реагування.

Для розроблення комп'ютерної програми прогнозування популярності текстової публікації instagram-блогу необхідно визначити інформаційно-структурні та лінгвістичні ознаки (параметри) тексту, які засвідчують його популярність і впливають на його популярність.

Найважливішими зовнішніми параметрами, що визначають популярність instagram-публікації, є кількість лайків та коментарів. Лайк – це функція, яку використовують для вираження позитивного ставлення користувачів до того чи іншого контенту [17]**.** Коментар – це запис під публікацією (постом), що дає можливість користувачеві виражати своє ставлення до розміщеного контенту [17].

Внутрішні ознаки текстотворення instagram-тексту є факторами впливу на популярність instagram-блогу. У роботі ставиться завдання визначити, які

параметри текстотворення необхідно враховувати блогеру для підвищення популярності блогу.

Запитання. Збільшити кількість коментарів під instagram-публікацією можна за допомогою використання запитань у тексті, адже це одна з відомих комунікативних стратегій організації діалогу (Рис. 1.1).

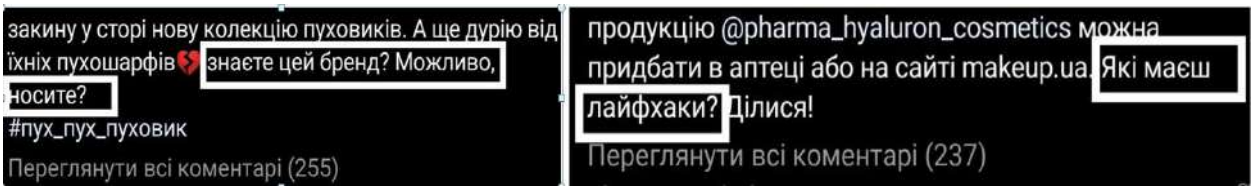


Рис. 1.1. Комунікативна стратегія запитань у публікації Instagram

Заголовок. На думку відомих seo-фахівців І. Ашманова та А. Іванова [1], суттєвим параметром, що впливає на популярність текстового контенту, є заголовок. Заголовком в Instagram називаємо вступну фразу, яка відділена від тексту відступом, піктограмою або відступом із крапкою (Рис. 1.2). «В інформаційно перенасиченому світі від заголовка великою мірою залежить, чи піде читач далі за текстом…» [11, 44].

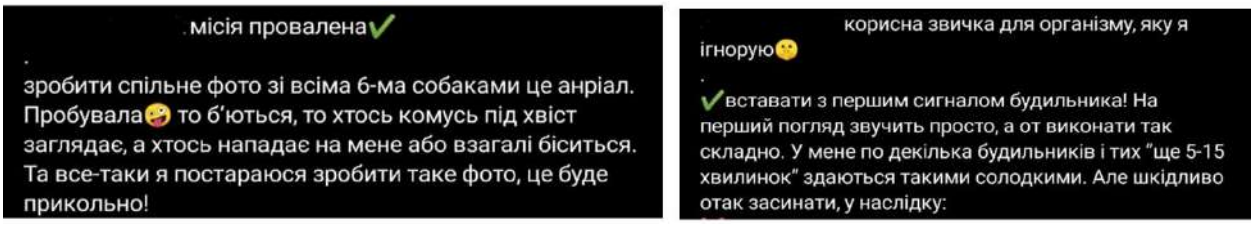


Рис. 1.2. Заголовки в Instagram-текстах

Хештеги. Особливими ознаками текстового контенту в Instagram є можливість використання хештегів (Рис. 1.3). Хештег – це ключове слово, фраза чи словосполучення, яке починається зі символа «#» і яке використовують у текстових описах публікації для тематичного об'єднання його з іншими дописами, що мають ідентичні хештеги [17]. Хештег є першим рівнем інформації для пошукового алгоритму та релевантного запиту. Хештег потрібен алгоритму пошуку для того, щоб:

- згрупувати пости за ключовими словами, темами;
- швидко знайти пост на потрібну тему;

- розширити аудиторію;

Рис. 1.3. Приклад хештегів в Instagram-тексті

Гіперпокликання. Характерною ознакою інтернет-комунікації є гіпертекстуальність – «можливість за допомогою гіперпокликань встановлювати миттєвий зв'язок між асоціативно залежними текстами, що розміщені на різних сайтах мережі» [15, 68].

Гіперпокликання – це «активний (виділений кольором) текст, зображення чи кнопка на веб-сторінці, натиснення на яку (активізація гіперпокликання) викликає перехід на іншу сторінку чи іншу частину поточної сторінки» [3]. Гіперпокликання в Instagram на інший обліковий запис створюється за допомогою символа «@» і має безанкорний принцип: є конкретним пошуковим запитом за адресою (Рис. 1.4).

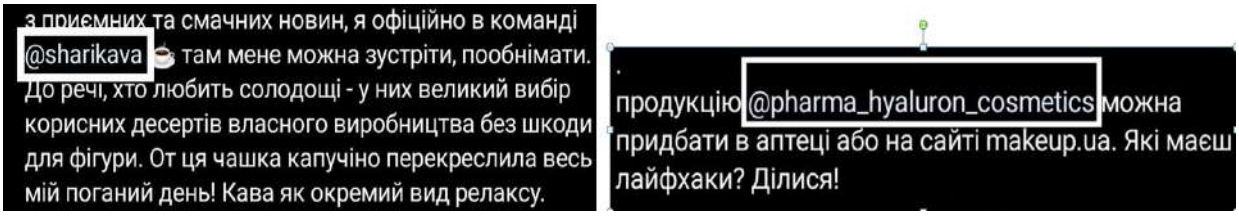


Рис. 1.4. Приклад гіперпокликань в Instagram-тексті

Використання хештегів та гіперпокликань в інтернет-тексті демонструє принцип використання seo-засобів комп'ютерного комунікативного каналу у ролі комунікативного коду текстотворення.

Піктограми. На сприйняття читачами текстового контенту в Instagram позитивно впливають піктограми, що урізноманітнюють публікацію та допомагають виразити емоції (Рис. 1.5).

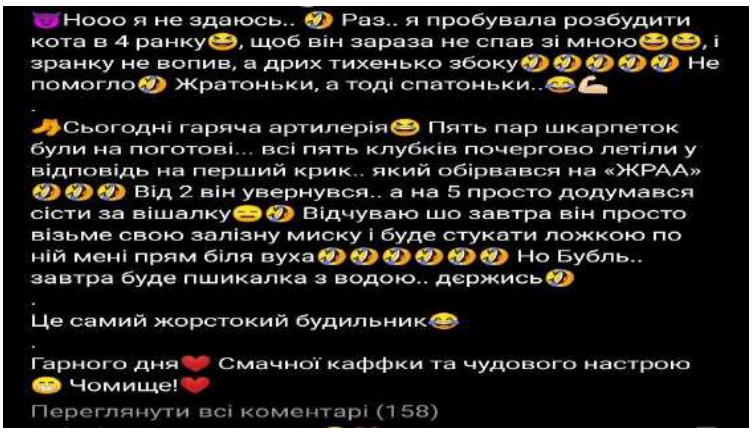


Рис. 1.5. Приклад піктограм в інтернет-тексті

Структурування тексту. Візуальне структурування тексту на абзаци за допомогою різних роздільників: відступів, пунктуаційних знаків, піктограм (Рис. 1.6).

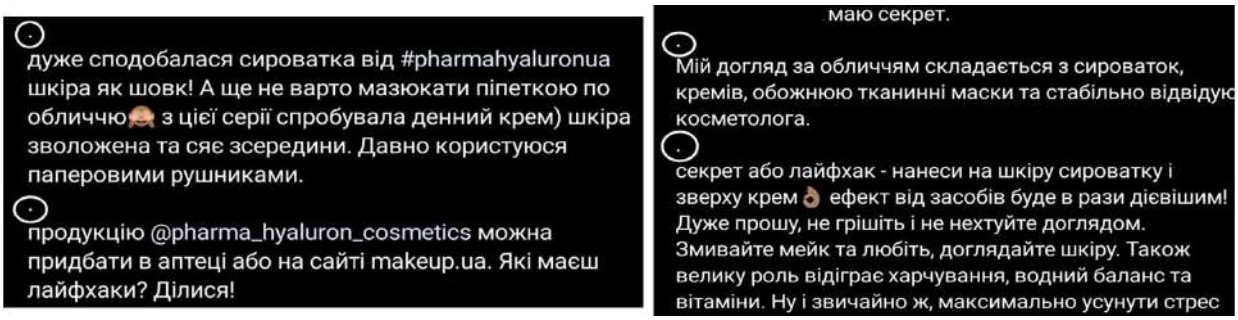


Рис. 1.6. Приклад структурованої Instagram-публікації

Обсяг текстової публікації. Обсяг тексту також має значення для популярності публікації. «І не лише з погляду читабельності для ока, а для алгоритмів-пошуковиків. Експерти називають різний оптимальний обсяг текстової публікації: від 200 до 1000 слів» [11, 45].

Тональність тексту. Крім того, на вподобання читачем текстової публікації впливає її тональність, а саме позитивне, негативне або нейтральне ставлення блогера до об’єктів, явищ, персоналій, згаданих у текстовій публікації.

Отже, для розроблення програми прогнозування популярності текстової публікації в Instagram було обрано такі досліджувані параметри: 1) кількість лайків instagram-публікації; 2) кількість коментарів під instagram-публікацією; 3) наявність заголовка; 4) кількість хештегів; 5) кількість гіперпокликань; 6)

обсяг текстової публікації; 7) кількість абзаців; 8) кількість роздільників між абзацами; 9) кількість піктограм; 10) кількість питальних речень; 8) тональність тексту.

Створення автоматичної системи прогнозування популярності текстових публікацій у соціальній мережі Instagram допоможе визначити, як впливають перераховані параметри на кількість читацьких вподобань текстової публікації.

РОЗДІЛ 2
СТВОРЕННЯ КОМП'ЮТЕРНОЇ ПРОГРАМИ АВТОМАТИЧНОГО АНАЛІЗУ ТЕКСТОВОЇ ПУБЛІКАЦІЇ У СОЦІАЛЬНІЙ МЕРЕЖІ INSTAGRAM

Розроблення програмного забезпечення автоматичної системи аналізу текстової публікації у соціальної мережі Instagram було поділено на декілька етапів:

- формування корпусу текстових публікацій мережі Instagram;
- створення комп'ютерної програми для автоматичного визначення й кількісного підрахунку лінгвістичних та інформаційно-структурних параметрів у текстовій публікації;
- створення комп'ютерної програми тонального аналізу текстової публікації;
- створення бази даних та автоматичний імпорт кількісних даних до бази даних;

2.1. Формування корпусу instagram-текстів

Першим завданням у створенні корпусу текстових публікацій мережі Instagram був пошуковий аналіз українськомовного облікового запису Instagram, читацька аудиторія якого становить близько 100 тисяч читачів, а кількість постів – більше двох тисяч. Також обов'язковою умовою добору текстового матеріалу дослідження була вимога, щоб обліковий запис не виступав спонсором матеріальних розіграшів у мережі Instagram, які підвищують активність блогу за рахунок збільшеної кількості лайків та коментарів. За визначеними вимогами було обрано обліковий запис @nata_fedorchuk, цільова аудиторя якого становить ≈ 99 200 читачів, а кількість постів – 3 268.

Для створення корпусу текстів був розроблений програмний код (Додаток А) для завантаження текстових публікацій за допомогою бібліотек requests [62] та lxml [56].

База даних корпусу текстів систематизує дані за 6-ма полями таблиці (Рис. 2.1):

1) порядковий номер;

2) дата текстової публікації;

3) автор публікації;

4) кількість лайків;

5) кількість коментарів;

6) текстова публікація.

ID	date_of_publication	auther	likes	comments	text
1	1 лютого	nata_fedorchuk	6570	62	корисна звичка для органі...
2	29 січня	nata_fedorchuk	4789	47	#slowlife то ніби про мене....
3	26 січня	nata_fedorchuk	6079	78	думками телепортуюся за...
4	22 січня	nata_fedorchuk	5504	70	тян Рін□|якось і не писал...
5	19 січня	nata_fedorchuk	3542	150	об'єм для тонкого волосся...
6	17 січня	nata_fedorchuk	3961	57	колаборація з армані та ві...
7	15 січня	nata_fedorchuk	7644	128	зробити спільне фото зі в...
8	14 січня	nata_fedorchuk	6022	129	нам потрібна твоя посміш...
9	13 січня	nata_fedorchuk	6979	92	сьогодні що якесь місячне...

Рис. 2.1. Фрагмент бази даних корпусу текстових публікацій

2.2. Автоматичне укладання бази даних внутрішніх параметрів інтернет-текстів

На другому етапі створення автоматичної системи прогнозування популярності текстової публікації необхідно було визначити у кожній текстовій публікації корпусу текстів внутрішні лінгвістичні та інформаційно-структурні параметри тексту й імпортувати кількісні дані з корпусу текстів про зовнішні параметри та отримані кількісні дані про внутрішні параметри до створеної за допомогою бібліотеки sqlite3 [64] бази даних «instagram_posts.db». Це завдання було реалізовано у послідовності здійснення таких операцій:

1. Автоматичне вилучення із бази даних корпусу текстів кожної текстової публікації, її порядкового номера, кількості лайків та коментарів.

2. Створення кодових скриптів для автоматичного визначення та підрахунку в кожній текстовій публікації внутрішніх інформаційно-структурних та лінгвістичних параметрів.

3. Виведення даних про кожну текстову публікацію за досліджуваними параметрами. На рис. 2.2 подано фрагмент параметризації однієї текстової публікації, де систематизовані дані про: порядковий номер публікації в корпусі текстів; кількість лайків; кількість коментарів; наявність заголовка (0 – якщо відсутній, 1 – якщо є); кількість хештегів; кількість гіперпокликань (підрахунок символів @); довжина текстової публікації (кількість всіх символів, включаючи піктограми та пробіли); кількість абзаців; кількість роздільників; кількість піктограм; кількість питальних речень.

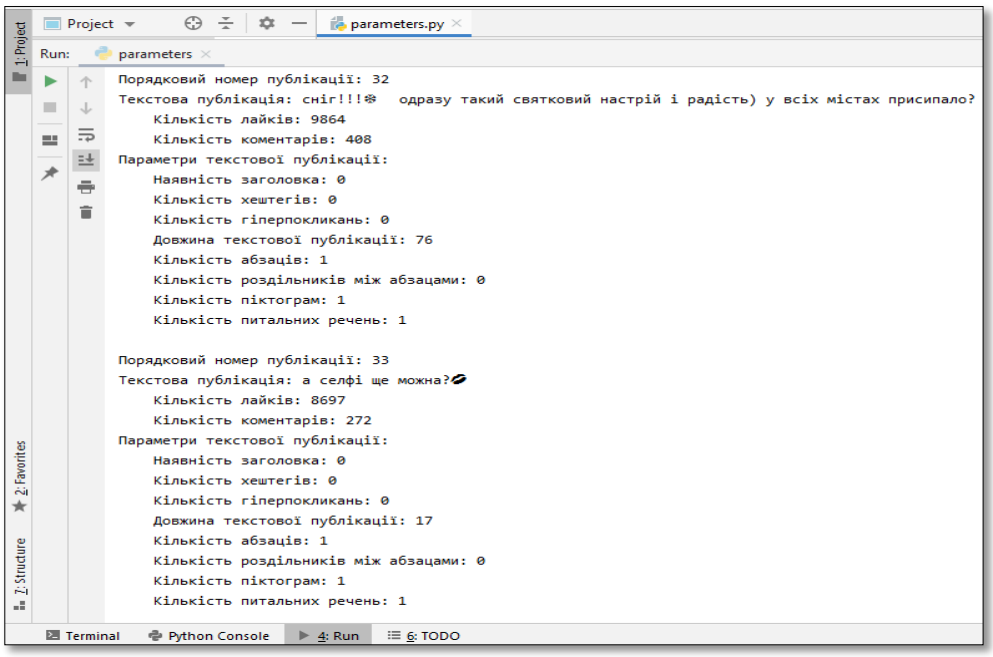


Рис. 2.2.Фрагмент виведення даних про параметризацію текстових публікацій

4. Створення структури бази даних «instagram_posts.db».

5. Імпорт даних, описаних у пункті 3, до бази даних «instagram_posts.db» (Рис. 2.4).

Розроблене програмне забезпечення (Додаток Б) автоматично вилучає кожну текстову публікацію, її порядковий номер, кількість лайків та коментарів із корпусу текстів, визначає лінгвістичні та інформаційно-структурні параметри

кожної текстової публікації за заданими маркерами й автоматично імпортує ці дані до бази даних «instagram_posts.db», яка буде використана для побудови автоматичної системи прогнозування популярності текстових публікацій.

2.3. Аналіз тональності instagram-текстів

Останнім параметром аналізу instagram-текстів є тональність тексту. Визначення цього параметра передбачає розроблення програмного забезпечення для визначення тональності кожного тексту та імпорту отриманих даних до бази даних «instagram_posts.db» (Рис. 2.4).

Тональний аналіз, або сентимент-аналіз, становить складний комплекс завдань, тому доречно розглянути розвиток цього вектора прикладної лінгвістики докладніше. Перші дослідження у галузі тонального аналізу тексту були проведені під час Другої світової війни (Р. Стагнер [42]). Комп'ютерний аналіз тональності тексту був започаткований Асоціацією комп'ютерної лінгвістики, заснованої в 1962 році, а «золотий стандарт» виявлення у тексті емоційно маркованих речень розроблено Дж. Вібе в 1999 році [48]. Із цього часу сентимент-аналіз став одним із найпопулярніших завдань у системах автоматичного оброблення природної мови, якому присвячено багато прикладних проектів відомих дослідників: В. Каспера [29], К. Єсенова [50], Т. Ваха [43], П. Бурнапа [45] та ін. .

Існує багато практичних застосувань тонального аналізу тексту, зокрема: у сфері бізнесу (дослідження Н. Каджі [27],Т. Вілсона [49]); у політиці (Б. О'Коннор [23], М. Томас [44], Т. Спренгер [39], Т. Маллен і Р. Малуф [36]); у створенні рекомендаційних систем (Л. Тервін, В. Хіл та інші [38], Б. Панг [37]); в інтернет-торгівлі (С. Вохра [47]); в урядовому регулюванні (Е. Спертус [41]) та в інших сферах життя суспільства.

Аналіз тональності тексту може проводитися:

1) на лексичному рівні: визначення відсотка позитивно, негативно та нейтрально забарвлених слів у тексті;

2) на синтаксичному рівні: тональність речення визначається сукупністю тональностей усіх його слів, а тональність тексту – сукупністю тональностей усіх речень.

На аналіз тональності тексту також впливають правила сполучуваності тональної лексики, фразеологізми, порядок слів у реченні, піктограми, пунктуація та інші випадкові чинники, які потрібно враховувати при визначенні тональності, зокрема:

- наявність заперечень та протиставлень у тексті;
- іронія та сарказм;
- неологізми;
- орфографічні помилки;
- випадки, коли слова в реченні, залежно від контексту, можуть набувати протилежного емоційного забарвлення (наприклад, речення *Ця книга має непередбачуваний сюжет* є позитивним, а речення *Ця людина має непередбачувану поведінку* – негативним);
- випадки, коли слова із позитивним чи негативним забарвленням набувають нейтральної конотації (наприклад, речення *Чи є продукція цієї компанії найкращою з-поміж усіх інших?* містить слово із позитивним забарвленням (*найкращою*), але за емоційним забарвленням це речення є нейтральним).

Більшість дослідників, серед яких і М. Лобур [34], визначають такі методи класифікації текстів за тональною ознакою: методи машинного навчання (Х. Кан та Д. Хан [26], Ч. Бхадане та Х. Далал [18]); методи на основі правил, що описані в роботах К. Манфреда [35], Д. Кана [28]); статистичні методи на основі тонального словника, представлені у роботі М. Табоади, Дж. Брук, М. Тофілоскі [32]).

У нашому дослідженні для аналізу тональності текстів соціальної мережі Instagram було розроблено програму за функціональним принципом роботи

інструмента для оброблення природних мов VADER [46], який було адаптовано на українськомовний текст.

VADER (ValenceAwareDictionaryandsEntimentReasoner) – це інструмент тонального аналізу текстів соціальних мереж для англійської мови. Він базується на тональному словнику слів та наборі правил, що впливають на визначення тональності всього тексту.

Розроблення комп'ютерної програми аналізу тональності текстів передбачало послідовне виконання таких завдань:

1. Створення корпусу текстових публікацій (див. § 2.1);

2. Автоматичне вилучення кожної текстової публікації та її порядкового номера із бази даних корпусу.

3. Попереднє оброблення текстових публікацій (токенізація за допомогою бібліотеки tokenize_uk [63]) та морфологічний аналіз (лематизація, частиномовна розмітка із застосуванням морфологічного аналізатора pymorphy2 [53]);

4. Визначення тональності слів з використанням українського тонального словника [9], [52] (приписування кожному слову позитивної (pos), негативної (neg) чи нейтральної (neu) тональності);

5. Розроблення правил сполучуваності тональної лексики для визначення тональності речення (із застосуванням бібліотек pymorphy2 [53], nltk [59]);

6. Розроблення правил оцінки впливу на тональність речення пунктуаційних знаків, слів та словосполучень у верхньому регістрі, піктограм у графічному та текстовому форматах (із застосуванням бібліотеки emoji[55]);

7. Виведення даних для кожного тексту (Рис. 2.3).

ID	pos	neg	neu
1	0.2	0.1	0.7
2	0.8	0	0.2
3	0	0.6	0.4
4	0	0	1
5	0.4	0.1	0.5
6	0.5	0	0.5
7	0.6	0.1	0.3
8	0.2	0	0.8
9	0.1	0.5	0.4
10	0.2	0.4	0.2

Рис. 2.3. Таблична систематизація тональних значень текстових публікацій.

Визначення тональності речення відбувається на основі підрахунку тональностей усіх слів з урахуванням правил сполучуваності тональної лексики, пунктуаційних знаків та піктограм. Результат – виведення трьох оцінок тональності речення (позитивне (pos), негативне (neg) та нейтральне (neu)). Їх сума не перевищує значення 1. Наприклад, для речення *Він добрий та працьовитий* програма приписує такі значення: позитивне – 0.531, нейтральне – 0.469, негативне – 0.0. Якщо до цього речення додати піктограму«☹», що виражає негативну емоцію, то значення позитивної оцінки тексту зменшиться до 0.347, нейтральної до 0.408, а негативне значення підвищиться до 0.245.

Розроблене програмне забезпечення автоматично вилучає кожну текстову публікацію та її порядковий номер із корпусу текстів, визначає тональність текстової публікації (цифрове значення позитивної, негативної та нейтральної тональності текстової публікації, що становить відсоток від 1 – рис. 2.3) та імпортує отримані дані (Додаток В) у базу даних «instagram_posts.db» (Рис. 2.4). Створена програма не враховує всіх аспектів, що впливають на оцінку тональності текстової публікації, тому що представляє лише перший етап розроблення автоматичної системи тонального аналізу українськомовних текстів соціальних мереж.

2.4. Структура бази даних «instagram_posts.db»

З метою створення автоматичної системи прогнозування популярності текстових публікацій в Instagram за послідовністю виконаних завдань, описаних у попередніх параграфах другого розділу, було автоматично сформовано реляційну базу даних «instagram_posts.db» (Рис. 2.4), що систематизує інформацію про 500 текстових публікацій одного блогера соціальної мережі Instagram: обліковий запис @nata_fedorchuk. Таблиця бази даних «instagram_posts.db» складається з 500-та рядків та 13-ти колонок:

1) порядковий номер публікації (ID);
2) кількість лайків (likes);
3) кількість коментарів (comments);
4) кількість хештегів (hastag);
5) кількість гіперпокликань на інші облікові записи Instagram (referense);
6) довжина текстової публікації (len);
7) кількість абзаців (paragraph);
8) кількість роздільників між абзацами (indent);
9) кількість піктограм (emoji);
10) кількість питальних речень (question);
11) кількісне значення позитивної тональності (pos);
12) кількісне значення негативної тональності (neg);
13) кількісне значення нейтральної тональності (neu).

ID	likes	com...	headl...	hastag	refer...	len	para...	indent	emoji	ques...	pos	neg	neu
1	6570	62	0	0	0	2	2	1	2	0	0.2	0.1	0.7
2	4789	47	0	1	1	0	1	0	1	0	0.8	0	0.2
3	6079	78	1	0	0	2	3	2	6	0	0	0.6	0.4
4	5504	70	0	1	0	1	1	0	1	0	0	0	1
5	3542	150	0	0	1	1	1	0	1	1	0.4	0.1	0.5
6	3961	57	0	1	1	0	1	0	1	0	0.5	0	0.5
7	7644	128	1	5	2	3	6	5	7	3	0.6	0.1	0.3
8	6022	129	1	1	0	3	5	4	9	1	0.2	0	0.8
9	6979	92	0	0	0	2	4	3	1	2	0.1	0.5	0.4
10	7218	91	1	0	1	2	3	2	3	0	0.2	0.4	0.2

Рис. 2.4. Фрагмент бази даних «instagram_posts.db»

РОЗДІЛ 3
СТВОРЕННЯ КОМПЮТЕРНОЇ ПРОГРАМИ АВТОМАТИЧНОГО ВИЗНАЧЕННЯ ФАКТОРІВ ВПЛИВУ НА ПОПУЛЯРНІСТЬ ТЕКСТОВОЇ INSTAGRAM-ПУБЛІКАЦІЇ

Створення моделі прогнозування популярності текстових публікацій у соціальній мережі Instagram можливе за допомогою методів машинного навчання. Машинне навчання – це метод застосування алгоритмів для автоматичного знаходження закономірностей в організації даних. Популярність методів машинного навчання при розв’язанні практичних задач пов’язана з можливістю аналізувати велику кількість даних з метою виявлення приховані закономірностей у структурі даних.

Для побудови програми прогнозування популярності текстової публікації застосовано модель множинної лінійної регресії та бібліотеки matplotlib [57], numpy [58], pandas [60] , mpl_toolkits [61] , sklearn [54].

3.1. Множинна лінійна регресія у побудові комп'ютерної моделі прогнозування

Процес побудови моделей прогнозування на основі взаємозв'язку між залежною змінною та незалежною (незалежними) змінними, що входять в рівняння регресії, називається регресійним моделюванням. Регресійний аналіз може бути використаний для встановлення причинно-наслідкових зв’язків між залежними і незалежними змінними. Змінні, які використовуються для пояснення інших змінних, називаються незалежними. Залежна змінна – це змінна, результат якої залежить від інших змінних [19].

Етапи регресійного моделювання:

1. Вибір моделі регресії.
2. Вибір об'єкта прогнозування – залежної змінної.
3. Вибір незалежних змінних, від яких залежить результат залежної змінної.
4. Укладання бази даних залежної та незалежних змінних.

5. Перевірка статистичної значущості моделі за статистичними коефіцієнтами.
6. Оцінка точності моделі.
7. Обчислення прогнозу на основі регресійного моделювання для прогнозування кількісного значення залежної змінної під впливом незалежних змінних.

Для побудови автоматичної системи прогнозування популярності текстової публікації обрано модель множинної лінійної регресії. Множинна лінійна регресія – це регресійна модель для встановлення взаємозалежності між залежною змінною (y) та двома або більше незалежними змінними (x), що представлена рівнянням:

$$y = b + m_1x_1 + m_2x_2 + \ldots + m_nx_n,$$

де:

- y – залежна змінна;
- x_1, x_2 , x_n– незалежні змінні;
- b – вимір зміни y щодо випадкових факторів, які не належать до регресійної моделі;
- m_1,m_2, m_n вимір зміни y у відношенні до x_1, x_2, x_n.

3.2. Автоматичне визначення прогностичних параметрів популярності текстової instagram-публікації

Створення програми прогнозування популярності текстових публікацій за допомогою моделі множинної лінійної регресії відбувається на основі бази даних «instagram_posts.db» (Рис. 2.4), параметрам якої надаються значення залежної та незалежних змінних:

- параметр «likes» (кількість лайків) обрано як залежну змінну (y);
- параметри: кількість коментарів (comments), кількість хештегів (hastag), кількість гіперпокликань (referense), довжина текстової публікації (len), кількість абзаців (paragraph), кількість роздільників між абзацами (indent), кількість піктограм(emoji), кількість питальних речень

(question), коефіцієнт позитивної тональності (pos), негативної тональності (neg) та нейтральної тональності (neu), – визначені незалежними змінними (*x*).

Як і в більшості алгоритмів машинного навчання, ми розділили дані на навчальний набір даних – 400 текстових публікацій (80%) та тестовий набір даних – 100 текстових публікацій (20%).

За навчальними даними бази даних «instagram_posts.db» (Рис.2.4.) будуємо графік (Рис. 3.1) розподілу фактичних і прогнозованих значень змінної «likes» за допомогою бібліотеки matplotlib [57]. На графіку представлений ймовірний розподіл залежної змінної «likes» з урахуванням усіх незалежних параметрів. На осі *yj* відкладено фактичні значення залежної змінної «likes». Кожне фактичне значення на площині представлене у вигляді точки блакитного кольору. На осі *yi* відкладено прогнозовані значення змінної «likes», що на площині представлене у вигляді точки синього кольору. Розподіл прогнозованих значень вважається точним, якщо він покриває розподіл фактичних значень.

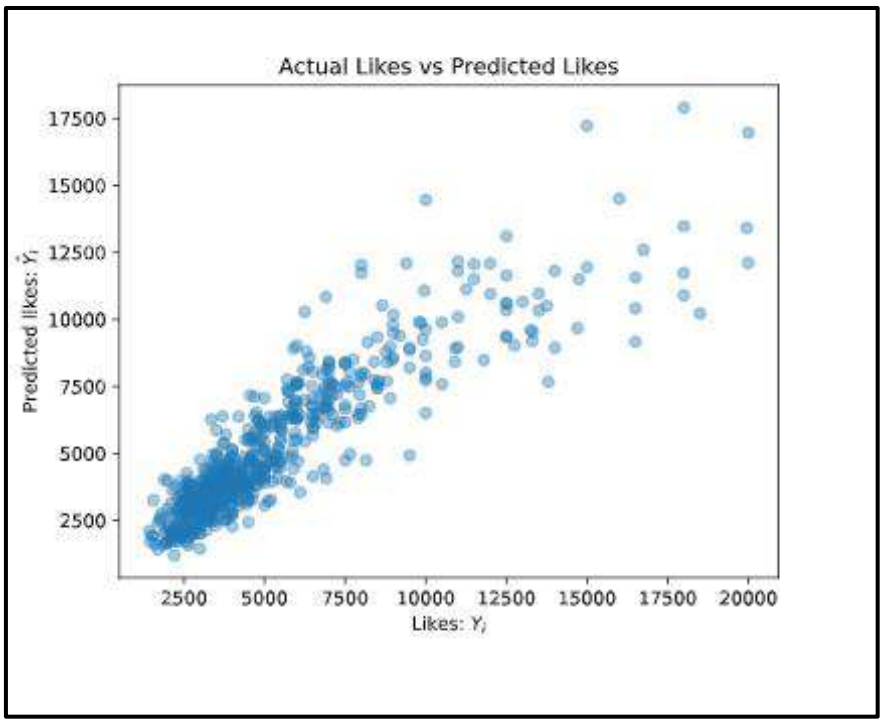


Рис. 3.1. Розподіл фактичних і прогнозованих значень залежної змінної «likes»

На рис. 3.1 показано, що розподіл прогнозованих значень залежної змінної відхиляється від розподілу її фактичних значень. Тому важливо з’ясувати ступінь залежності змінної «likes» від незалежних змінних. Найпростішим критерієм, що визначає ступінь залежності між двома показниками є коефіцієнт кореляції. У множинній лінійній регресії незалежні змінні або мають позитивне лінійне відношення до залежної змінної, або негативне лінійне відношення, або взагалі не мають відношення. Від'ємне лінійне співвідношення означає, що зі збільшенням значень *x* значення *y* зменшуються. Аналогічно, позитивна лінійна залежність означає, що зі збільшенням значень *x* значення *y* також збільшуватимуться. Коефіцієнти кореляції допомагають визначити, яка незалежна змінна має більшу вагу, більший вплив на залежну змінну. Наприклад, коефіцієнт 0.254 вплине на залежну змінну менше ніж 0.765.

Було створено програмне забезпечення для автоматичного обчислення коефіцієнта кореляції (Додаток Г). Кореляційний аналіз допоміг виділити незалежні змінні, що найбільше впливають на розподіл прогнозованих значень змінної «likes». Ці дані представлено у таблиці 3.1.

Таблиця 3.1. Кореляції змінних

Незалежні змінні	**Залежна змінна «likes»**
comments	0.507
headline	**0.830**
hastag	**0.775**
reference	0.226
len	**0.698**
haragraph	0.158
indent	0.158
emoji	0.249
question	0.191
pos	**0.892**
neg	**0.904**
neu	**0.635**

Найвищий коефіцієнт кореляції мають параметри: присутність заголовка (headline), кількість хештегів (hastag), довжина тексту (len), позитивна тональність (pos), негативна тональність (neg) та нейтральна тональність (neu).

Для оцінки точності множинної лінійної регресійної моделі було розроблено програмне забезпечення для обчислення коефіцієнта детермінації R^2 (Додаток Ґ). Коефіцієнт R^2 визначається за формулою 1- $\frac{U}{v}$, де: U – сума квадратів відхилення лінії регресії від фактичних значень, а v – сума квадратів відхилень лінії регресії від середнього значення. Чим ближче значення коефіцієнта детермінації до одиниці, тим краще модель описує статистичні дані. Наприклад, якщо намагаємося передбачити популярність текстової публікації за змінною «likes» через наявність у текстовій публікації незалежних змінних заголовка (headline) та довжини тексту (len), а R^2 для нашої моделі дорівнює 0,72, то це означає, що незалежні змінні «headline» і «len» разом пояснюють 72% змін залежної змінної «likes». Якщо додати до моделі незалежну змінну негативної тональності тексту (neg), очевидно, що зміниться значення R^2. Припустимо, що нове значення R^2 становить 0,95. Це означає, що незалежні змінні «headline», «len» і «neg» разом пояснюють 95% коливань залежної змінної «likes». Коефіцієнт детермінації вважається задовільним, якщо його критичне значення становить: $R^2 \geq 0{,}70$.

Модель вважається точною, якщо фактичні значення максимально наближені до прогнозованих. Для нашого набору даних R^2 навчальних даних (trainscore) наближена до R^2 тестових даних (testscore) (Рис. 3.2), тому робимо висновок про достатню точність створеної моделі.

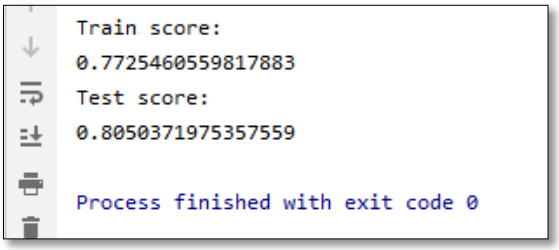


Рис. 3.2. Оцінка точності прогнозування

Для покращення моделі вилучимо з вихідного файлу бази даних «instagram_posts.db» незалежні змінні, які, за результатами кореляційного аналізу, не впливають на залежну змінну «likes»: кількість коментарів (comments), кількість гіперпокликань (referense), кількість абзаців (paragraph), кількість роздільників між абзацами (indent), кількість піктограм (emoji) та кількість питальних речень (question), (Рис. 3.3).

ID	likes	headline	hastag	len	pos	neg	neu
1	6570	0	0	2	0.2	0.1	0.7
2	4789	0	1	0	0.8	0	0.2
3	6079	1	0	2	0	0.6	0.4
4	5504	0	1	1	0	0	1
5	3542	0	0	1	0.4	0.1	0.5
6	3961	0	1	0	0.5	0	0.5
7	7644	1	5	3	0.6	0.1	0.3
8	6022	1	1	3	0.2	0	0.8
9	6979	0	0	2	0.1	0.5	0.4
10	7218	1	0	2	0.2	0.4	0.2

Рис. 3.3. Фрагмент бази даних «instagram_posts.db» після вилучення незалежних змінних з низьким коефіцієнтом кореляції

Вилучення незалежних змінних призвело до суттєвого покращення моделі прогнозування. На рис. 3.4 представлена діаграма розподілу фактичних і прогнозованих значень залежної змінної «likes» після вилучення незалежних змінних з низьким коефіцієнтом кореляції.

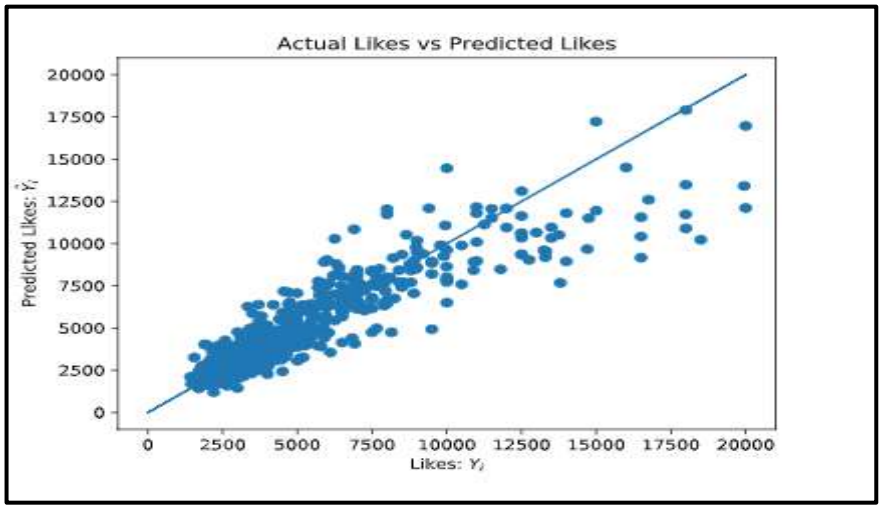


Рис.3.4. Діаграма розподілу фактичних і прогнозованих значень змінної «likes» після покращення моделі

Для перевірки точності створеної програми прогнозування було здійснено перевірку тестового набору текстових публікації. Порівнявши прогнозоване моделлю значення із фактичним значенням змінної «likes», ми з'ясували, що точність прогнозування коливається від 87% до 94%. Це свідчить про точність побудованої моделі прогнозування популярності текстової публікації (Рис. 3.5).

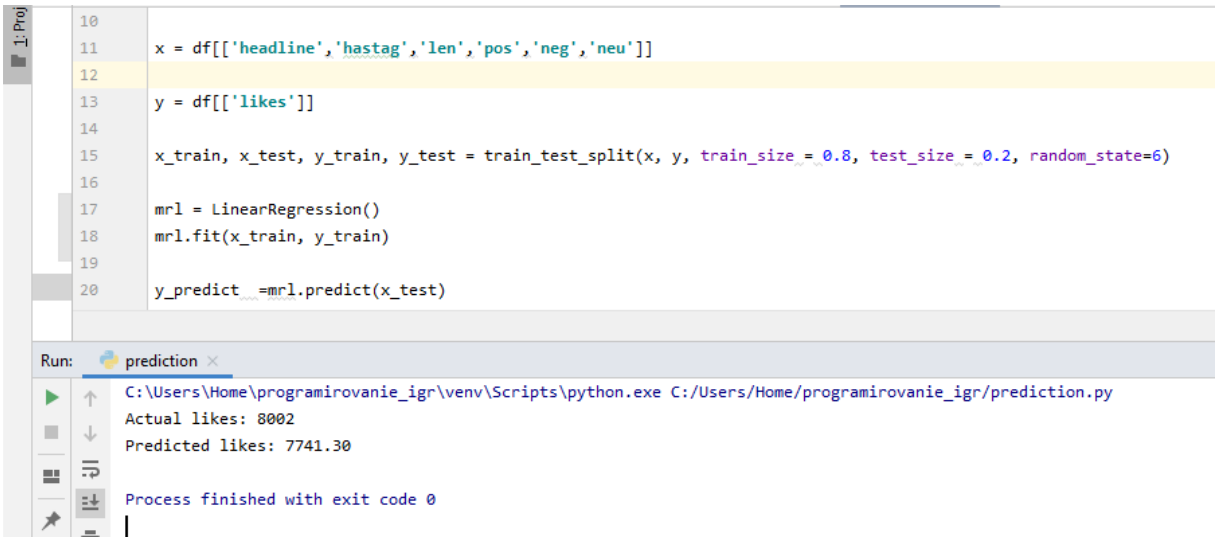


Рис. 3.5. Приклад виведення фактичного та прогнозованого значення змінної «likes»

ВИСНОВКИ

Проведене дослідження має міждисциплінарний прикладний характер: воно ґрунтоване на методологічних засадах інтернет-лінгвістики, SEO-копірайтингу, математичної та комп'ютерної лінгвістики. Синтез міждисциплінарних методів, застосованих у вирішенні прикладного лінгвістичного завдання, демонструє одну з головних тенденцій розвитку сучасної лінгвістики – інтеграцію з іншими науками. Актуальність, наукова та практична значущість дослідження визначаються не тільки адаптацією нелінгвістичних методів у дослідженні мовного об'єкта, а також метою та самим об'єктом дослідження.

Створена автоматична система визначення факторів впливу на популярність текстової публікації у соціальній мережі Instagram – перша аналітична система в українському прикладному мовознавстві, спрямована на аналіз instagram-тексту. У процесі створення системи було проведено аналіз інформаційного каналу та знакового коду instagram-комунікації і визначено особливості організації полілогового instagram-дискурсу, які можна прокласифікувати за такими ознаками: 1) зовнішні фактори визначення популярності instagram-тексту: лайки та коментарі; 2) внутрішні фактори впливу на популярність instagram-тексту, які поділяються на: а) структурно-інформаційні: хештеги та гіперпокликання; б) лінгвістичні: запитання, заголовок, абзаци, обсяг тексту, тональність тексту; в) піктограмні.

Вплив визначених факторів на популярність (кількість лайків) instagram-тексту було встановлено на основі прогностичної математичної моделі множинної лінійної регресії, використаної у машинному навчанні. Ступінь залежності кількості лайків від прогностичних факторів популярності був визначений за допомогою автоматичного обчислення коефіцієнта кореляції, який показує, що фактичний вплив на популярність instagram-тексту мають параметри: присутність заголовка, кількість хештегів, довжина тексту, позитивна, негативна та нейтральна тональність. Значення коефіцієнта кореляції у розподілі трьох оцінок тональності демонструє вищий ступінь

залежності змінної (кількість лайків) від тональної оцінки «позитивний текст» (0.892) та «негативний текст» (0.904), натомість коефіцієнт кореляції емоційно нейтрального тексту становить 0.635. Тому можна зробити висновок, що на популярність instagram-тексту впливає тільки позитивна та негативна тональність. Таким чином, з 11 прогностичних факторів (незалежних змінних) лише 5 (присутність заголовка, кількість хештегів, довжина тексту, позитивна та негативна тональність) мають фактичний вплив на популярність instagram-тексту, визначену за залежною змінною (кількість лайків).

Створена автоматична система апробована лише на 500-ах публікаціях одного облікового запису @nata_fedorchuk із соціальної мережі Instagram та за однією залежною змінною (кількість лайків). Достовірність отриманих результатів буде значно вищою, якщо цю систему застосувати для аналізу декількох облікових записів, а також використати в прогностичній математичній моделі множинної лінійної регресії ще одну залежну змінну – кількість коментарів. Крім того, необхідно проаналізувати коефіцієнт кореляції між залежною змінною «кількість коментарів» та незалежною змінною «кількість запитань», які, на нашу думку, повинні мати високе значення коефіцієнта кореляції. Ці завдання становлять перспективу нашого дослідження.

Результати роботи автоматичної системи визначення факторів впливу на популярність текстової публікації у соціальній мережі Instagram демонструють високий ступінь точність прогнозування на основі моделі множинної лінійної регресійної: коефіцієнт детермінації R^2, встановлений для навчальних даних (0.773) наближений до R^2 тестових даних (0.805).

Отримані результати машинного навчання визначили перспективні завдання у подальшій роботі над створенням автоматичної системи визначення факторів впливу на популярність текстової публікації і можливість представити цю систему як інформаційний лінгвістичний продукт у формі веб-сайту мережі Інтернет для широкого кола користувачів-блогерів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1.Ашманов И., Иванов. А. Оптимизация и продвижение сайтов в поисковых системах. Изд. 3-е. Питер, 2011. 464 с.

2.Бергельсон М. Б. Языковые аспекты виртуальной коммуникации. *Вестник МГУ. Серия 19. Лингвистика и межкультурная коммуникация.* 2002. № 1. С. 55–67.

3. Гіперпосилання. URL: https://uk.wikipedia.org/w/index.php? (дата звернення: 12.02.2020).

4. Гончарова Е. А. Медиальный аспект модуса формулирования текста как проблема стилистики. Вып. 7. Белград, 2008. С. 11–20.

5. Иванова Т. С. Речевое поведение интернет-общения. *Вестник АГУ.* № 3. 2011. С. 81-85.

6. Кирющенко В. В. Язык и знак в прагматизме. Санкт-Петербург: ЕУСПб, 2008. 200 с.

7. Компанцева Л. Ф. Интернет-комммуникация: когнитивно прагматический и лингвокультурологичкий аспекты. Луганск: Знание, 2007. 444 с.

8. Окландер М. А. Комплекс інтеренет-комунікацій у маркетингу. *Маркетинг в Україні*. 2008. № 3. С. 29 – 35.

9. Романишин М., Романюк А. Тональний словник української мови на основі сентимент-анотованого корпусу. *Українське мовознавство*. 2013. № 43. С. 6374.

10. Савицька Н. Л. Маркетинг у соціальних мережах: стратегії та інструменти на ринку B2C. Маркетинг і цифрові технології. 2017. № 1, С. 20 –33. URL: https://mdt-opu.com.ua/index.php/mdt/article/view/5 (дата звернення: 08.02.2020).

11. Чекмишев О. В., Ярошенко Л. А. Основи якісного блогерства. Київ. 2014. 48 с.

12. Чернявская В. Е. Лингвистика текста. Поликодовость. Интертекстуальность. Интердис-курсивность. Москва. 2009. 248 с.

13. Чернявская В. Е. Текст в медиальном пространстве. Москва: Флинта. 2013. 224 с.

14. Чичкань Г. Мова та стиль інтернет-комунікації. URL: http://journal47.at.ua/publ/chichkan_galina_mova_ta_stil_internet_komunikaciji/1-1-0-5 (дата звернення: 05.02.2020).

15. Шилінська І. Лінгвостилістичні аспекти інтер-комунікації. *Мандрівець*. 2014. № 5. С. 67–69.

16. Фатурова В. М. Інтернет-середовище як фактор психологічного розвитку комунікативного потенціалу особистості: автореф. дис. ... канд. психол. наук: 19.00.07. Київ, 2004. 27 с.

17. Barabasi A. L. Linked: The new science of networks. Cambridge: *Perseus Publishing*. 2002. 229 p.

18. Bhadane C., Dalal H., Doshi H. Sentiment analysis: Measuring opinions. International Conference on Advanced Computing Technologies and Applications. Procedia Computer Science. 2015. Vol., No. 45. P. 808 – 814.

19. Cohen J., Cohen P., West S. G, Aiken L. S. Applied multiple regression/cor relation analysis for the behavioral sciences, 3rd Ed. Mahwah. NJ: Lawrence Erlbaum Associates. 2003.

20. Crystal D. Language and the Internet. 2-d edition. Cambridge : Cambridge University Press, 2006. 33 p.

21. Denysyuk. Z. Internet communication as a trend of everyday social practices. *Culture and Modernity*. No.1. P. 27-31.

22. Fix U. Die Ästhetisierung des Alltags – am Beispiel seiner Texte. Zeitschrift für Germanistik. 2001. H 11.1. P. 36-53.

23. O'Connor B., Balasubramanyan R., Routledge B., Smith N. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. Pittsburgh:

Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. 2010. P. 237–251.

24. Goddin S. Tribes: We Need You to Lead Us. New York: Portfolio. 2008.

25. Grimov O. A. Social and cultural practices of the individual in social networks. Extended abstract of candidate's thesis. Kursk. 2014.

26. Kang H., Yoo S. J. Han D. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. Expert Systems with Applications. Vol. 39, Is. 5. 2012. P. 6000–6010.

27. Kaji N., Kitsuregawa M. Building lexicon for sentiment analysis from massive collection of html documents. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2007. P. 301–324.

28. Kan D. Rule-based approach to sentiment analysis at ROMIP 2011. URL: http://www.slideshare.net/dmitrykan/rulebased-approach-to-sentiment-analysis-at-romip-2011 (Last accessed: 11.02.2020).

29. Kasper W., M. Vela. Sentiment Analysis for Hotel Reviews. Proceedings of the Computational Linguistics-Applications Conference. Poland: Polskie Towarzystwo Informatyczne, Katowice. 2011. P. 45–52.

30. Katz E., Lazarsfeld P. Personalinfluence: The part played by people in the flow of mass communications. Piscataway: Transaction Publishers. 2005. 400 p.

31. Khurshid A. Affective Computing and Sentiment Analysis: Metaphor, Ontology, Affect and Terminology. Berlin: Springer Science & Business Media. 2011. 164 p.

32. Lexicon-based methods for sentiment analysis. *Computational Linguistics*. Vol. 37. 2011. P. 267-307.

33. Liu B., Dekker M. Sentiment Analysis and Subjectivity. New York: Handbook of Natural Language Processing. 2010. 192 p.

34. Lobur M., Romaniuk A., Romanyshyn M. Defining an approach for deep sentiment analysis of reviews in Ukrainian. Lviv Polytechnic National University Computer-Aided Design Department. 2012. P. 37–66.

35. Manfred K., Fahrni A., Petrakis S. PolArt: A robust tool for sentiment analysis. In Proceedings of the 17th Nordic Conference of Computational Linguistics. Vol. 4. 2009. P. 235-238.

36. Mullen T., Malouf R. Taking sides: User classification for informal online political discourse. InternetResearch, California. 2008. 190 p.

37. Pang B., Lee L., Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing. Vol. 10. 2002. P. 79–86.

38. PHOAKS: A system for sharing recommendations. In Communications of the Association for Computing Machinery (CACM). 2007. P. 59–62.

39. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. A. Tumasjan, T. Sprenger, P. Sandner, I. Welpe. 2010. P. 189–224.

40. Schurina V. Classification of comic speech genres communicative Internet space. *Izvestyia VHPU*. Vol. 2. 2014. P. 39–43.

41. Spertus E. Smokey: Automatic recognition of hostile message. In Proceedings of Innovative Applications of Artificial Intelligence. 2013. P. 1058–1065.

42. Stagner R. The cross-out technique as a method in public opinion analysis. *The Journal of Social Psychology*. Vol. 11, No. 1, 1940. P. 79 –90.

43. Text mining for market prediction: A systematic review / ed. T. Wah, A. Nassirtoussi, S. Aghabozorgi, and D. Ngo. Expert Systems with Applications. Vol. 41, No. 16. 2014. P. 7653–7670.

44. Thomas M. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts/ ed. M. Thomas, B. Pang, and L. Lee. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2006. P. 327–335.

45. Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack/ ed. P. Burnap et al – Social Network Analysis and Mining. Vol. 4, No. 1. 2016. P. 1–14.

46. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text / ed. C. Hutto, and E. Gilbert. ICWSM. The AAAI Press. 2014.

47. Vohra S., Teraiya J. Applications and Challenges for Sentiment Analysis: A Survey. *Int. Journal of Engineering Research & Technology*. 2013. P. 1-5.

48. Wiebe J., Bruce R., O'Hara T. Development and use of a gold-standard data set for subjectivity classifications. 1999. P. 246–253.

49. Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. 2009. P. 399–433.

50. Yessenov K., Misailovic S. Sentiment Analysis of Movie Review Comments. Massachusetts Institute of Technology. 2009. P. 121–142.

51. Zajtseva S. The Internet-communication as a new form of interindividual communication. URL: http://ukrmova.com.ua/zmist-zhurnalu/vipusk-11/internet-spilkuvannyayak-nova-forma-mizhosobistisno%D1%97-komunikaci%D1%97/ (Last accessed: 12.02.2020).

СПИСОК ДЖЕРЕЛ ПРОГРАМНИХ БІБЛІОТЕК

52. Український тональний словник. URL: https://github.com/lang-uk/tone-dict-uk (дата звернення: 04.01.2020).

53. Морфологический анализатор pymorphy2. URL: https://pymorphy2.readthedocs.io/en/latest/ (дата звернення: 05.01.2020).

54. A set of python modules for machine learning and data mining. URL: https://pypi.org/project/sklearn/ (Last accessed: 12.01.2020).

55. Emoji for Python. URL: https://pypi.org/project/emoji/ (Last accessed: 05.01.2020).

56. Lxml XML and HTML with Python. URL: https://lxml.de (Last accessed: 24.12.2019).

57. Matplotlib. URL: https://matplotlib.org/ (Last accessed: 16.01.2020).

58. NumPy. URL: https://pypi.org/project/numpy/ (Last accessed: 12.01.2020).

59. Natural Language Toolkit. URL: http://www.nltk.org/ (Last accessed: 16.01.2020).

60. Pandas. URL: https://pypi.org/project/pandas/ (Last accessed: 13.01.2020).

61. Python plotting package. URL: https://pypi.org/project/mpl/ (Last accessed: 14.01.2020).

62. HTTP for Humans. URL: http://docs.python-requests.org/en/master/ (Last accessed: 24.12.2019).

63. Simple python lib to tokenize texts tokenize_uk. URLhttps://github.com/lang-uk/tokenize-uk (Last accessed: 25.12.2019).

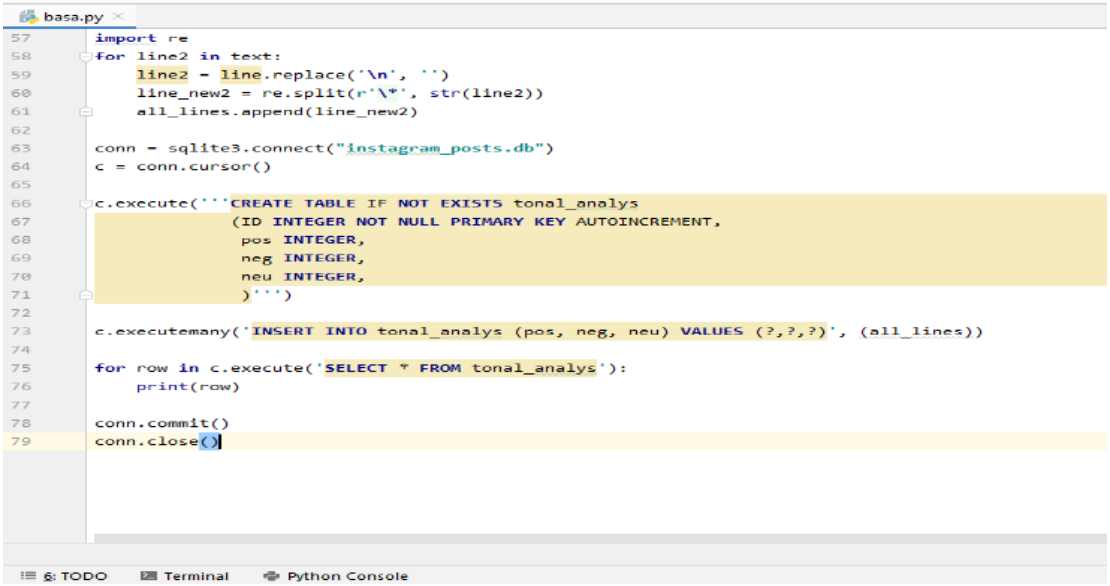
64. What Is SQLite? URL: https://www.sqlite.org/index.html (Last accessed: 13.01.2020).

ДОДАТКИ

Додаток А. Фрагмент програмного коду формування корпусу текстових публікацій

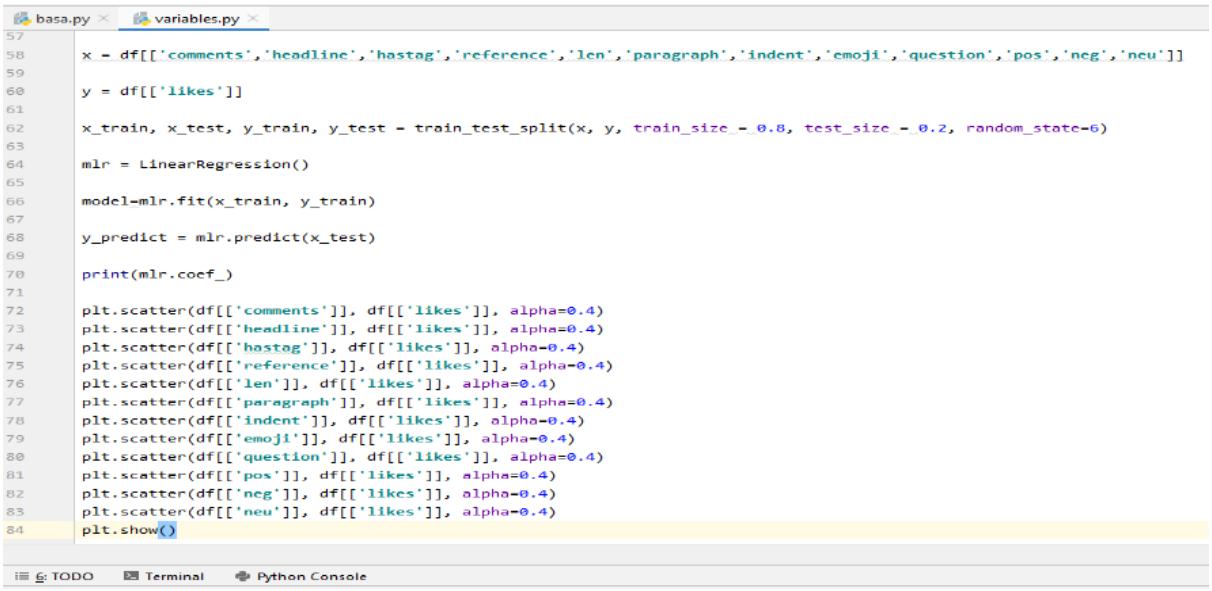
Додаток Б. Фрагмент програмного коду формування бази даних «instagram.posts.db» та внесення параметрів.

Додаток В. Фрагмент програмного коду імпорту результатів аналізу тональності текстових публікацій у базу даних «instagram_posts.db»



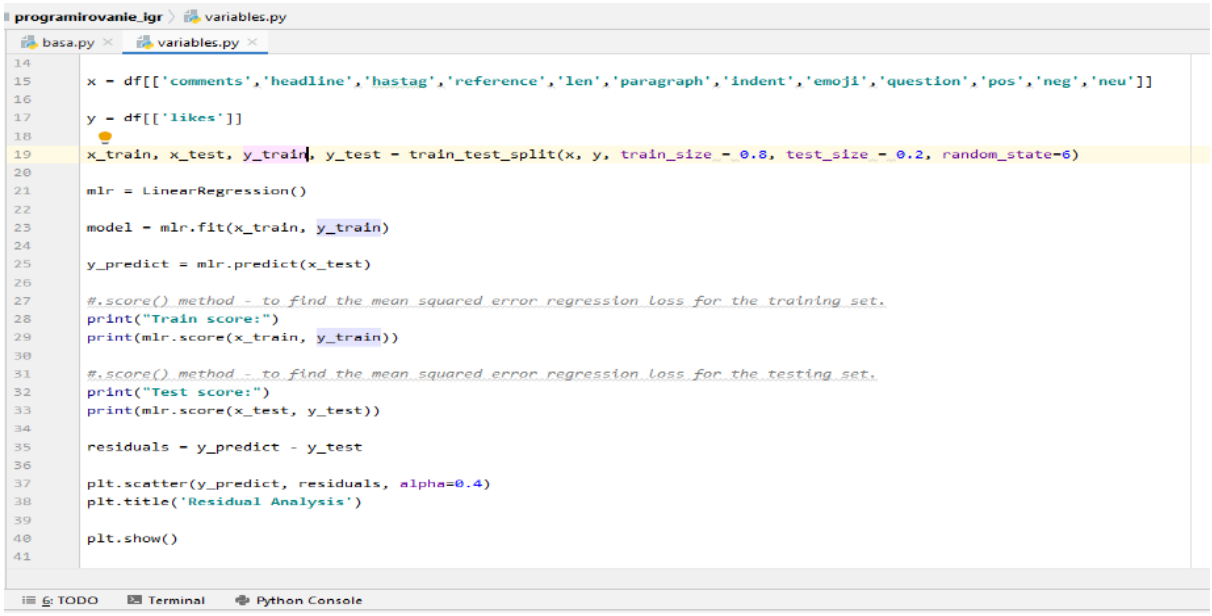
```
basa.py
import re
for line2 in text:
    line2 = line.replace('\n', '')
    line_new2 = re.split(r'\*', str(line2))
    all_lines.append(line_new2)

conn = sqlite3.connect("instagram_posts.db")
c = conn.cursor()

c.execute('''CREATE TABLE IF NOT EXISTS tonal_analys
             (ID INTEGER NOT NULL PRIMARY KEY AUTOINCREMENT,
              pos INTEGER,
              neg INTEGER,
              neu INTEGER,
              )''')

c.executemany('INSERT INTO tonal_analys (pos, neg, neu) VALUES (?,?,?)', (all_lines))

for row in c.execute('SELECT * FROM tonal_analys'):
    print(row)

conn.commit()
conn.close()

6: TODO  Terminal  Python Console
```

Додаток Г. Фрагмент програмного коду для визначення коефіцієнтів кореляції між залежною змінною та незалежними змінними.



```
basa.py  variables.py

x = df[['comments','headline','hastag','reference','len','paragraph','indent','emoji','question','pos','neg','neu']]

y = df[['likes']]

x_train, x_test, y_train, y_test = train_test_split(x, y, train_size = 0.8, test_size = 0.2, random_state=6)

mlr = LinearRegression()

model=mlr.fit(x_train, y_train)

y_predict = mlr.predict(x_test)

print(mlr.coef_)

plt.scatter(df[['comments']], df[['likes']], alpha=0.4)
plt.scatter(df[['headline']], df[['likes']], alpha=0.4)
plt.scatter(df[['hastag']], df[['likes']], alpha=0.4)
plt.scatter(df[['reference']], df[['likes']], alpha=0.4)
plt.scatter(df[['len']], df[['likes']], alpha=0.4)
plt.scatter(df[['paragraph']], df[['likes']], alpha=0.4)
plt.scatter(df[['indent']], df[['likes']], alpha=0.4)
plt.scatter(df[['emoji']], df[['likes']], alpha=0.4)
plt.scatter(df[['question']], df[['likes']], alpha=0.4)
plt.scatter(df[['pos']], df[['likes']], alpha=0.4)
plt.scatter(df[['neg']], df[['likes']], alpha=0.4)
plt.scatter(df[['neu']], df[['likes']], alpha=0.4)
plt.show()

6: TODO  Terminal  Python Console
```

Додаток Ґ. Фрагмент програмного коду для визначення коефіцієнта детермінації.



АНОТАЦІЯ

Назва конкурсної роботи: «Застосування моделі множинної лінійної регресії у прогнозуванні популярності українськомовного instagram-тексту».

Актуальність дослідження: Короткі тексти соціальних мереж, зокрема instagram-тексти, обмежені обсягом 2200 символів, можуть аналізуватися лише за зовнішніми факторами пошукової оптимізації, тому що статистичні методи визначення ключових слів та метод визначення семантичного ядра тесту, які використовуються сьогодні у пошуковій оптимізації, не є ефективними для аналізу коротких текстів. На сьогодні ще не існує систем для комплексного SEO-аналізу, мінімальних за обсягом, текстових публікацій у популярній соціальній мережі Instagram. Вивчення текстових особливостей цього дискурсу, а також текстових і позатекстових факторів, що впливають на популярність instagram-блогу – актуальне завдання сучасної комп'ютерної лінгвістики.

Мета дослідження – створити автоматичну систему визначення факторів впливу на популярність текстової публікації у соціальній мережі Instagram.

Досягнення мети передбачає виконання таких **завдань:**

1) сформувати корпус текстових instagram-публікації;

2) визначити лінгвістичні та структурні параметри, які зумовлюють популярність текстової публікації;

3) розробити програмне забезпечення для автоматичного визначення у текстах лінгвістичних та інформаційних параметрів, які можуть мати вплив на популярність текстової публікації;

4) розробити комп'ютерну програму автоматичного визначення прогностичних параметрів популярності текстової instagram-публікації на основі моделі множинної лінійної регресії.

Об’єктом дослідження є блогова українськомовна instagram-публікація.

Предметом дослідження є інформаційні та лінгвістичні фактори популярності українськомовного instagram-тексту.

Методи дослідження: методи статистичного аналізу, метод машинного навчання, метод моделювання множинної лінійної регресії, методика проведення тонального аналізу тексту.

Інформаційна база дослідження: мова програмування Python та її бібліотеки: emoji [55], lxml [56], matplotlib [57], mpl_toolkits [61], nltk [59], numpy [58], pandas [60], pymorphy2 [53], requests [62], sklearn [54], sqlite3 [64], tokenize_uk [63].

Матеріал дослідження: 500 текстових полілогів облікового запису @nata_fedorchuk із соціальної мережі Instagram.

Загальна характеристика роботи: Конкурсна робота складається зі вступу, трьох розділів, висновків, списку використаної літератури (51 позиція), списку джерел програмних бібліотек (13 позицій) та додатків (5 фрагментів програмного коду).

У **Вступі** розкрито актуальність теми дослідження, сформульовано мету та завдання, визначено об'єкт, предмет, методи дослідження, інформаційну базу та матеріал дослідження.

У першому розділі **«Лінгвістичні ознаки та інформаційні особливості інтернет-тексту»** проаналізовано основні інформаційно-структурні (зовнішні й внутрішні) та лінгвістичні ознаки instagram-тексту, які визначаються прогностичними факторами популярності instagram-публікації.

Другий розділ **«Створення комп'ютерної програми автоматичного аналізу текстової публікації у соціальній мережі Instagram»** представляє опис розроблення програмного забезпечення для автоматичного визначення прогностичних факторів популярності instagram-публікації та автоматичного укладання двох баз даних: бази даних корпусу instagram-текстів та бази даних кількісних характеристик прогностичних факторів популярності instagram-текстів.

Третій розділ **«Створення комп'ютерної програми визначення факторів впливу на популярність текстової instagram-публікації»**

присвячений опису машинного навчання, організованого за математичною моделлю множинної лінійної регресії, яка покладена в основу концепції створення комп'ютерної програми визначення факторів впливу на популярність текстової instagram-публікації. Також у цьому розділі обчислено статистичні характеристики:

1) коефіцієнт кореляції для встановлення ступеня залежності між чинниками впливу та показником популярності (кількість лайків) instagram-публікації;

2) коефіцієнт детермінації, що дає оцінку точності прогнозування, проведеного за моделлю множинної лінійної регресії.

У текстах трьох розділів конкурсної роботи подано 15 рисунків, які ілюструють особливості організації структури instagram-тексту та демонструють результати роботи створених комп'ютерних програм.

У **Висновках** зроблено аналіз отриманих результатів дослідження та визначено перспективи подальшої роботи над створенням автоматичної системи визначення факторів впливу на популярність текстової instagram-публікаці й представленням її як інформаційного продукту в мережі Інтернет.