Віддалений доступ до інформації через онлайн-портали. Створення онлайн-порталів для клієнтів, де вони можуть переглядати медичну інформацію про своїх тварин, отримувати результати аналізів та звертатися до лікаря віддалено, забезпечує зручність і доступність інформації для власників тварин.

Використання технологій штучного інтелекту (AI) та машинного навчання (ML) у ветеринарії. Не менш важливим є використання технологій штучного інтелекту й машинного навчання, яке може значно поліпшити діагностику та лікування тварин. Алгоритми машинного навчання можуть аналізувати великі обсяги клінічних даних та допомагати ветеринаріям у точній діагностиці різних захворювань.

Отже, сучасні методи й технології обробки даних ветеринарних клінік відіграють важливу роль у поліпшенні обслуговування клієнтів та наданні якісної ветеринарної допомоги. Електронні системи, IoT, технології AI та інші інноваційні рішення сприяють оптимізації робочих процесів та покращенню якості надання ветеринарних послуг.

Список використаних джерел

1. Mitchell R. Web Scraping with Python: Collecting More Data from the Modern Web. 2nd Edition. O'Reilly, 2018. 290 p.

2. Heydt M. Python Web Scraping Cookbook: Over 90 proven recipes to get you scraping with Python, micro services, Docker and AWS. Packt Publishing, 2018. 364 p.

3. Beautiful Soup Documentation. URL: https://www.crummy.com/software/ BeautifulSoup/ bs4/doc/

УДК 004.912

Бездушний В. О., здобувач 2 курсу спеціальності 122 Комп'ютерні науки,

Штовба С. Д., д-р техн. наук, професор, професор кафедри інформаційних технологій

ВИЯВЛЕННЯ ПРИХОВАНОЇ ЛАЙКИ В ТЕКСТОВИХ ПОВІДОМЛЕННЯХ ЗА АНАЛІЗОМ ВІЗУАЛЬНО ПОДІБНИХ СИМВОЛІВ

Донецький національний університет імені Василя Стуса, м. Вінниця

Приховані лайливі слова – це слова, що зазнали певних замін символів, але під час їх читання пересічний користувач легко розуміє їх лайливий сенс. Такі заміни роблять навмисно, щоб обійти автоматичні фільтри повідомлень у чатах, коментарях тощо. Для прикладу можна навести таку просту заміну

fool на *f00l*. У початковому слові замінено літери *oo* на нулі – *00*. Слова залишилися візуально дуже схожими, тому і *fool*, і *f00l* сприймається як лайливі з тотожними значеннями. Автоматичні фільтри, які базуються на посимвольному порівнянні слів, слово *fool* ідентифікують як лайливе, а слово *f00l* – ні, бо воно відсутнє у словниках образливої лексики.

Нами пропонується підхід до виявлення прихованих лайливих слів на основі аналізу візуально схожих символів. Якщо деяке слово вдається заміною одного чи кількох символів на візуально схожі символи перетворити на лайливе, тоді вважаємо, що приховане лайливе слово виявлено. Щоб реалізувати цю ідею, потрібно мати списки візуально подібних символів. Нами знайдено кілька матриць сплутувань для деякого набору літер англійського алфавіту, зокрема [1–4]. Ми агрегували матриці з цих робіт та доповнили їх відсутніми коефіцієнтами сплутувань для букв верхнього й нижнього регістрів та коефіцієнтами сплутувань цифр з буквами. Після цього виділили перелік найбільш схожих пар символів. Ці пари схожих символів і будемо аналізувати для виявлення спотворених лайливих слів. Фрагмент переліку найбільш схожих пар символів наведено нижче:

1 → l, i, I;
4 → a, A;
7 → t, T;
0 → o, u, O, U;
5 → S;
8 → B;
9 → g, q, R;
c → e;
L → I;
z → s;
G → O, C.

Також ми сформували словник базових лайливих слів, щоб було на що спиратись під час пошуку прихованої лайки. У базовий словник увійшло 169 слів.

Аналіз тексту відбувається у 3 етапи. Спочатку аналізоване речення розбивається на токени – окремі слова. Далі кожен токен направляється на грубу перевірку на схожість з словами із базового словника за відстанню Левенштейна. Якщо кількість елементарних перетворень аналізованого слова у базове лайливе слово не перевищує порогове значення, тоді запускається процедура тонкої перевірки. Тонка перевірка здійснюється, щоб встановити, чи можна перетворити аналізоване слово на лайливе заміною візуально схожих символів. Тонка перевірка здійснюється за бектрекінг-алгоритмом, який синтезує нові слова, враховуючи кожен символ із вхідного слова та потенційні заміни цих

символів на візуально схожі. Якщо синтез виявляється успішним, тоді аналізоване слово вважається приховано лайливим, і алгоритм повертає відповідне слово із базового словника.

Проілюструємо роботу алгоритму на такому реченні: *The ztubid fuekJng infant bItcb ass fock crawled into my ass the stupid fucking dense grass*. Це речення містить 8 лайливих слів, 4 із яких словникові і 4 – спотворені. Результат аналізу цього речення за запропонованим алгоритмом наведено на рис. 1. Із нього видно, що виявлено усі 8 лайливих слів. Лайливі слова зі словника було виявлено відразу без запуску бектрекінг-алгоритму. Тривалість аналізу такого речення у розробленому вебзастосунку становить 127 мілісекунд.

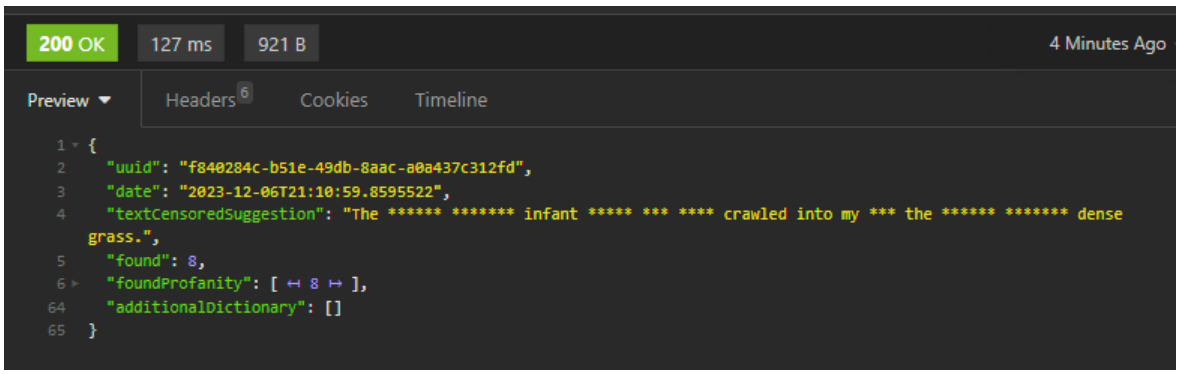

Рисунок 1. Результат аналізу тестового речення

Для порівняння з конкурентами згенеровано 3 такі тестові речення, що містять 6, 8 та 8 лайливих слів:

(1) *A fuck1ng man out a focklnq blue sweater fat ass sat at zh1tt1ng the st0p1d fuckinb desk*;

(2) *The ztubid fuekJng infant bItcb ass fock crawled into my ass the stupid fucking dense grass*;

(3) *Dude tried to fuekinq impress her ugly bitch face and fuekinb failed because the stubJd fucking c00n of is an Idjot.*

Результати тестування (табл. 1) свідчать, що запропонований алгоритм значно переважає конкурентів.

Таблиця 1 – Порівняння з конкурентами

Сервіс	**Кількість виявлених лайливих слів**			**Рівень виявлення лайливих слів**
	речення (1)	**речення (2)**	**речення (3)**	
Readable.com	4	4	3	50 %
Sightengine	3	5	2	45,5 %
Webpurify	3	4	3	45,5 %
PurgoMalum	3	3	3	41 %
Запропонований	5	6	7	82 %

Запропонований алгоритм протестовано на розміченій базі коментарів з *Kaggle Toxic Comment Classification Challenge* (табл. 2). Перевірка здійснювалася для коротких коментарів довжиною до 500 символів. Із таблиці видно, що запропонований алгоритм рідко коли виявляє лайливе слово в нейтральних коментарях – коментарях третьої групи. Образливі та погрозливі коментарі – коментарі з другої групи у 75,6 % випадків містять лайливе слово, яке виявляється запропонованим алгоритмом. Більшість токсичних коментарів першої групи також містить правильні або спотворені лайливі слова. Отже, виявлені за запропонованим алгоритмом приховані лайливі слова можуть розглядатися як додатковий інформативний атрибут для створення моделей виявлення токсичного контенту в соціальних мережах. Принциповою відмінністю запропонованого алгоритму є те, що він базується на аналітичному підході, а не на індуктивному. Тому він здатен виявляти не лише відомі спотворені лайливі слова, але і нові, які будуть створені на основі нових варіантів заміни символів.

Таблиця 2 – Результати тестування на датасеті Kaggle Toxic Comment

Група	**Тип коментаря**	**Кількість коментарів**	**Виявлено лайливе слово**	**Не виявлено лайливого слова**
1	Toxic	9 812	5 975 61 %	3 837 39 %
	Severe toxic			
2	Obscene	6 822	5 156 75,6 %	1 666 24,4 %
	Threat			
	Insult			
	Identity hate			
3	Neutral	50 030	1 604 3,2 %	48 426 96,8 %

Список використаних джерел

1. Loomis, J. M. (1982). Analysis of tactile and visual confusion matrices. *Perception & Psychophysics, 31,* 41–52.

2. Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics, 22,* 487–490.

3. Townsend, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics, 9,* 40–50.

4. Dunn-Rankin, P., Leton, D. A., Shelton, V. F. (1968). Congruency factors related to visual confusion of English letters. *Perceptual and Motor Skills, 26(2),* 659–666.