УДК 004.056

РОЗПІЗНАВАННЯ ШАХРАЙСЬКИХ ОГОЛОШЕНЬ ПРО ОРЕНДУ ЖИТЛА ЗА ДОПОМОГОЮ ОБРОБКИ ПРИРОДНОЇ МОВИ МЕТОДАМИ МАШИННОГО НАВЧАННЯ З ВИКОРИСТАННЯМ СИНТЕТИЧНИХ ДАНИХ

С. Ю. Дрозд[1,a], І. В. Стьопочкіна[1]

[1] Навчально-науковий Фізико-технічний інститут

Анотація

У даній роботі проводиться дослідження з метою розпізнавання шахрайських оголошень про оренду житла за допомогою обробки природної мови та методів машинного навчання. Для експериментів було використано набір даних, розміщених на міжнародній платформі Craigslist, що містить правдиві та фішингові оголошення. Для вирішення проблеми дисбалансу датасету було згенеровано додаткові синтетичні зразки шахрайських оголошень за допомогою рекурентної нейронної мережі LSTM та мовної моделі ChatGPT. Крім того, над наборами даних з синтетичними зразками та окремо без синтетичних зразків було застосовано алгоритми балансування (випадкова надмірна вибірка, SMOTE-NC, випадкова недостатня вибірка) та класове зважування. За даними кожного з вихідних наборів та для новоутворених збалансованих наборів було побудовано класифікаційні моделі, на основі алгоритмів k-найближчих сусідів (k-nearest neighbor metho, KNN), наївного Байєсівського класифікатора (Multinomial Naive Bayes Classifier, MNB), дерева рішень (Decision Tree Classifier, DTC), випадкового лісу (Random Forest Classifier, RFC), методу опорних векторів (Support Vector Machine, SVM), логістичної регресії (Logistic Regression, LR), багатошарового персептрона (Multilayer Perceptron Classifier, MLP). За результатами аналізу метрик точності (roc-auc, precision, recall, f1-score) визначено, що використання синтетичних зразків суттєво покращує якість класифікації. Найефективнішим методом балансування виявився SMOTE-NC, а найкращим алгоритмом класифікації – MLPClassifier. Розроблена класифікаційна модель може допомогти користувачам ефективно розпізнавати оголошення про оренду з ознаками шахрайства і зменшити ризик потенційної шахрайської діяльності.

Ключові слова: Фішинг, оренда, обробка природної мови, машинне навчання, синтетичні дані

Вступ

У зв'язку із початком війни з Росією 24 лютого 2022, тисячі українців були вимушені покинути свої домівки в пошуках безпечного місця перебування. За статистикою, кількість переселенців у межах України становить 4 893 079 осіб, а закордон виїхало 7 млн 996 тисяч біженців [1].

Хоча Україна та інші європейські держави намагаються забезпечити всіх переселенців житлом, через великий потік міграції значна частина біженців змушені шукати притулок самостійно. Більшість розглядають оголошення про оренду в Інтернеті, що створює сприятливі умови для поширення інтернет-шахрайства. Загалом, інтернет-шахрайство (або так званий «фішинг») розвивається в ритм з розвитком інформаційних технологій. Якщо раніше фішинг асоціювався перш за все з розсилкою шкідливих електронних листів та спаму, то зараз шахраї урізноманітнили техніки обману та розширили горизонти свого впливу на інші сфери. Так, наприклад, шахрайство все частіше зустрічається в галузі онлайн-оренди. Зловмисники не лише заманюють потенційних жертв на свої шахрайські сайти, але й проникають на всесвітньовідомі веб-платформи, такі як Booking, Craigslist чи OLX. Зокрема, як повідомляють працівники інтернет-ресурсу OLX, в Україні після початку війни у розділі "OLX Нерухомість"було виявлено порушення в 7% оголошень від загальної кількості контенту, серед яких 654 публікацій – з підозрою у шахрайстві [2].

Часто зловмисникам вдається обманути жертву до того, як їх фіктивні оголошення будуть ідентифіковані як спам та видалені. Біженці натрапляють в інтернеті на житло з привабливою ціною, і, не перевіривши достовірність оголошення, відразу пересилають аванс. Проте прибуваючи за вказаною адресою, жертви обману не знаходять ні самого житла, ні того, хто обіцяв їх поселити.

Тому проблема шахрайства в сфері оренди житла є гострою і актуальною. Наразі існує багато методів

[a]sofi.drozd.13@@gmail.com

боротьби з фішингом. Більшість основані на виявленні спамових електронних листів та розпізнаванні шахрайських веб-сайтів за допомогою так званих традиційних методів. Сюди належать метод білих і чорних списків, пошук аномалій, евристичний та сигнатурний аналізи, тощо. Проте оскільки зловмисники постійно удосконалюють способи обману, традиційні засоби для розпізнавання фішингу не працюють достатньо ефективно. Виникає потреба в залученні нових, інноваційних методів для протидії інтернет-шахрайству.

Відносно новою технологією для виявлення інтернет-шахрайства з-поміж правдивого контенту є застосування машинного навчання та, передусім, обробка природної мови. Аналіз природної мови та машинне навчання успішно впроваджують для виявлення фейкових новин [3], телекомунікаційного [4] та фінансового [5] шахрайства, та в багатьох інших галузях [6]. Прикладом, що ілюструє ефективність застосування даної методики в сфері виявлення фальшивих оголошень, є дослідження вчених з США та Пакистану «Виявлення підроблених оголошень про роботу за допомогою підходів машинного навчання та обробки природної мови» [7]. У ході експерименту вдалося досягти 99.9% точності у розпізнаванні фіктивних оголошень.

У цьому дослідженні застосовуємо метод аналізу природної мови з використанням методів машинного навчання для виявлення шахрайських оголошень про оренду житла, розміщених на офіційних платформах для оренди. Головною метою роботи є встановити доцільність застосування даного методу для розпізнавання шахрайського контенту та збудувати класифікатор для ефективного вивлення фіктивних рентних оголошень. У ході роботи з метою визначення найефективнішої стратегії для створення найкращого класифікатора буде проведено кілька експериментів з проведенням аналізу різних класифікаційних моделей, навчених на основі збалансованої та незбалансованої вибірки, з використанням та без використання синтетичних зразків даних.

1. Дані та матеріали

Для проведення дослідження було використано набір даних, відкритий для публічного доступу у репозиторії github [8], що містить записи про оренду житла на платформі Craigslist -– ресурсі для розміщення різного типу електронних оголошень, яким активно послугуються жителі США та користувачі ще з 69 інших країн світу.

Набір даних складається з 2732 оголошень про оренду, категоризованих як «спам» і «не спам» протягом місяця, шляхом перевірки первинно завантажених повідомлень на ідентифікацію їх як спамових модераторами Craigslist та видалення з сайту. Перевірки відбувалися через 7, 14 та 30 днів після завантаження. У підсумку було визначено 21 шахрайське оголошення.

Під час перевірки набору даних виявлено, що записи містять дублікати, а реальна кількість унікальних оголошень становить 2487 (2469 правдивих та 18 шахрайських). Очевидно, що дані сильно дисбалансові, що може спричини проблеми для результативного навчання класифікатора. Для вирішення проблем дисбалансу, було прийнято рішення розширити міноритарний (менший) клас за рахунок створення додаткових синтетичних зразків шахрайських оголошень за допомогою розширеної нейронної мережі LSTM та новітньої мовної моделі чатботу ChatGPT.

У результаті, було додатково згенеровано 75 шахрайських оголошень, серед яких 11 створених нейронної мереженою та 64 - моделлю ChatGPT. У підсумку ми отримали 2023 реальних та 93 шахрайських оголошення (співвідношення класів шахрайських та правдивих оголошень 1:22 проти 1:137 до очистки та додавання синтетичних зразків відповідно), що зменшило, але не вирішило проблему дисбалансу.

Наступним кроком у даній роботі було балансування наборів даних з синтетичними та без синтетичних зразків, окремо застосовуючи над кожним з цих вихідних датасетів алгоритми випадкової надмірної вибірки, випадкової недостатньої вибірки, SMOTE-NC та класового зважування.

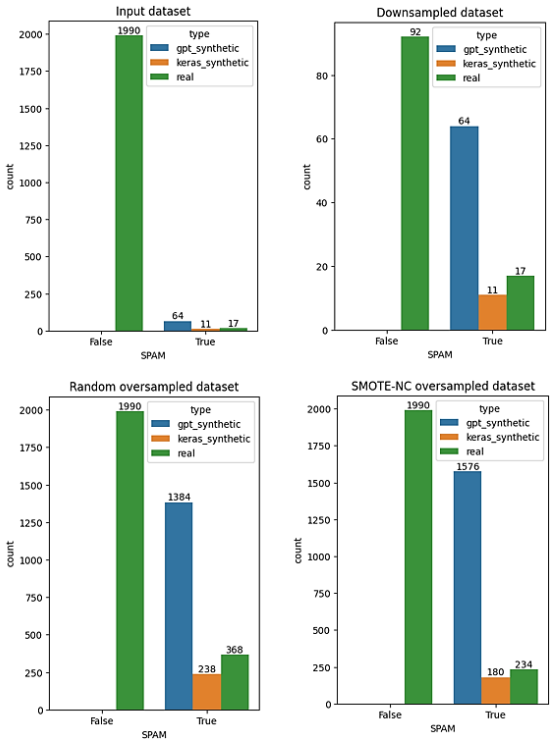

Рис. 1. Набори даних, збалансовані за допомогою різних алгоритмів

Таким чином було отримано 10 датасетів (2 вихідних – з синтетичними та без синтетичних зразків, для кожного з вихідних наборів - по 3 датасети, збалансованих різними алгоритмами та по 1 датасету із застосуванням класового зважування). Отримані набори даних (рис. 1) були використані для навчання моделей та проаналізовані для вибору найкращої комбінації даних, методу балансування та алгори-

тму класифікації для створення найоптимальнішої та найточнішої моделі для розпізнавання шахрайських оголошень про оренду.

Для навчання і валідації класифікаторів дані були поділені у співвідношенні 60% до 40% відповідно з розмежуванням навчальних і валідаційних зразків та із застосування стратифікації за класом і природою даних (реальні, згенеровані нейронною мережею LSTM, отримані з ChatGPT).

2. Методи дослідження

2.1. Генерація синтетичних даних

Для збільшення потужності множини шахрайських оголошень було застосовано дві методики для генерації синтетичних прикладів – розширену нейронну мережу LSTM та мовну модель ChatGPT. LSTM (Long short-term memory) -– це рекурентна нейронна мережа (RNN), що здатна обробляти послідовності даних (мову, відео, тощо). Ця характеристика робить дану мережу ідеальним інструментом для обробки та прогнозування даних. У даному досліджені було збудовано модель (рис. 2), що має два приховані шари LSTM з 256 одиницями пам'яті, щільний шар з функцією активації softmax для виведення прогнозів, та шар Dropout з параметром 0.2 для уникнення перенавчання. Модель отримує послідовний вхід з 3D-форми та компілюється з функцією втрати categorical_crossentropy і оптимізатором Adam.

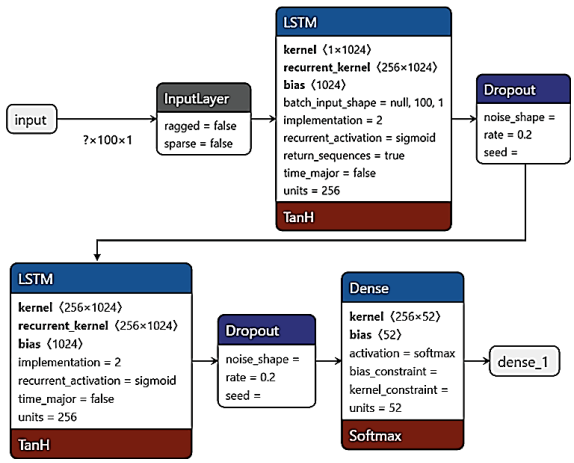

Рис. 2. Модель нейронної мережі LSTM для генерації синтетичних оголошень

На основі 18 наявних реальних шахрайських оголошень, за допомогою даної моделі було згенеровано ще 11 зразків. Оскільки для розширення набору даних цієї кількості недостатньо, також було застосовано інноваційну мовну модель –- чат-бот ChatGPT. Моделі було надано для прикладу реальні шахрайські оголошення, за якими ChatGPT попросили згенерувати подібні зразки. Як результат, чат-бот створив ще 64 унікальних синтетичних шахрайських оголошення.

2.2. Балансування даних та класове зважування

Крім генерації синтетичних зразків, для вирішення проблеми дисбалансу було вирішено застосувати кілька алгоритмів балансування та класове зважування з перспективою подальшого аналізу точностей класифікаторів, навчених на наборах даних, оброблених цими алгоритмами, для вибору найоптимальнішого з них. Загалом, для балансування було використано такі алгоритми:

1. Випадкова надмірна вибірка – просте випадкове дублюванням екземплярів міноритарного класу (шахрайських оголошень)
2. Synthetic Minority Over-sampling Technique for Nominal and Continuous (SMOTE-NC) – забезпечує генерацію синтетичних зразків мінорного класу, шляхом вибору декількох його найближчих сусідів і створення нових зразків на основі їхніх характеристик.
3. Випадкова недостатня вибірка – просте випадкове видалення екземплярів мажоритарного класу (правдивих оголошень) до отримання однакової кількості екземплярів обох класів.
4. Класове зважування – надання ваги кожному класу під час обчислення функції втрат таким чином, щоб вплив кожного класу на загальну функцію втрат був однаковим, використовуючи наступну формулу (1):

$$w_j = \frac{\mathrm{n}_{\mathrm{samples}}}{\mathrm{n}_{\mathrm{classes}} \cdot \mathrm{n}_{\mathrm{samples}_{ji}}} \tag{1}$$

де w_j – вага для кожного класу (j означає клас), $\mathrm{n}_{\mathrm{samples}}$ – це загальна кількість записів у наборі даних, $\mathrm{n}_{\mathrm{classes}}$ – загальна кількість класів, $\mathrm{n}_{\mathrm{samples}_{ji}}$ – загальна кількість рядків відповідного класу.

Алгоритми боротьби з дисбалансом були окремо застосовані для набору даних без синтетичних зразків та з синтетичними зразками.

2.3. Класифікація оголошень

Для розпізнавання фішингових оголошень на основі 10 отриманих наборів даних було навчено 7 алгоритмів класифікації – метод k-найближчих сусідів (k-nearest neighbor metho, KNN), наївний Байєсівський класифікатор (Multinomial Naive Bayes Classifier, MNB), дерево рішень (Decision Tree Classifier, DTC), випадковий ліс (Random Forest Classifier, RFC), метод опорних векторів (Support Vector Machine, SVM), логістична регресія (Logistic Regression, LR), багатошаровий персептрон (Multilayer Perceptron Classifier, MLP). Таким чином всього було отримано 70 моделей. Для визначення найкращої моделі в задачі розпізнавання шахрайських оголошень використовували наступні метрики для вимірювання точності:

$$precision = \frac{TP + FP}{TP} \tag{2}$$

$$recall = \frac{TP + FN}{TP} \tag{3}$$

$$f1\_score = 2 * \frac{precision + recall}{precision * recall} \quad (4)$$

$$TPR = \frac{TP}{TP + FN} \quad (5)$$

$$FPR = \frac{FP}{FP * TN} \quad (6)$$

$$ROC - AUC = \int TPR(FPR) dFPR \quad (7)$$

де TP – кількість правильно передбачених позитивних прикладів, FP – кількість неправильно передбачених позитивних прикладів, FN – кількість неправильно передбачених негативних прикладів.

При аналізі метрик, при порівнянні precision та recall перевагу надавали recall, оскільки в даному досліджені головним завданням є знайди і правильно класифікувати якогомога більше шахрайських оголошень.

3. Результати

За результатами аналізу точностей класифікаторів, побудованих за незбалансованими наборами даних, що містили лише реальні записи (таб. 1), та з синтетичними зразками (таб. 2), було встановлено, що додавання синтетичних прикладів суттєво впливає на покращення точності класифікації.

Таблиця 1. Метрики точностей моделей, побудованих на незбалансованих даних без синтетичних зраків

Model	roc-auc	precision	recall	f1-score
DTC	0,59	0,23	0,19	0,19
KNC	0,67	0,23	0,03	0,06
LR	0,9	0,17	0,03	0,06
MLP	0,92	0,1	0,01	0,02
MNB	0,76	0	0	0
RFC	0,87	0	0	0
SVC	0,54	0	0	0

Таблиця 2. Метрики точностей моделей, побудованих на незбалансованих даних з синтетичними зразками

Model	roc-auc	precision	recall	f1-score
DTC	0,8	0,71	0,6	0,64
KNC	0,89	0,76	0,59	0,66
LR	0,95	0,85	0,52	0,65
MLP	1	0,94	0,78	0,85
MNB	0,9	0,6	0,01	0,08
RFC	1	1	0,58	0,74
SVC	0,86	0,74	0,29	0,39

Зокрема, можна спостерігати ріст TPR (рис. 3) для кожного класифікаційного алгоритму (за винятком multinomial Naive Bayes classifier (MNB), що мав низькі показники f1-score 0 та 0.08 для наборів

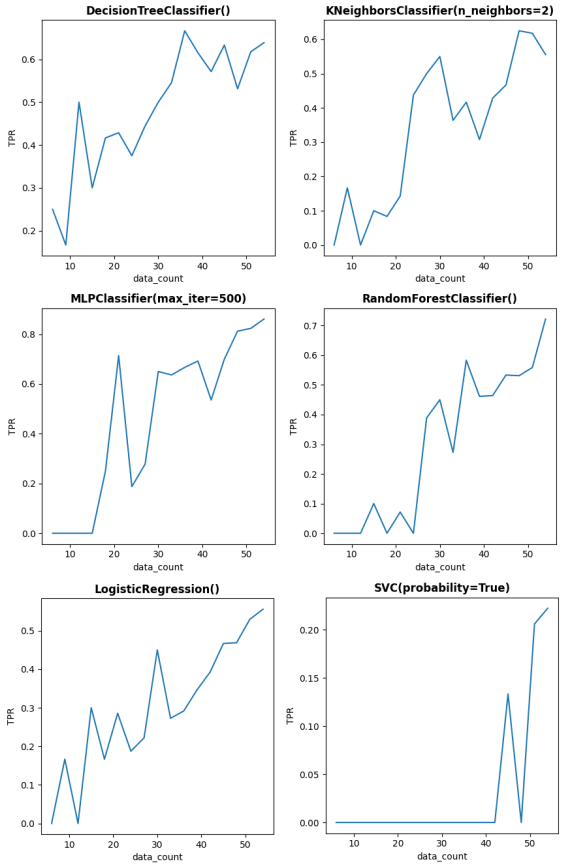

Рис. 3. Залежність TPR від кількості шахрайських оголошень у навчальному наборі на тестовому наборі даних

даних без та з синтетичними прикладами відповідно).

При цьому на незбалансованому наборі загальна точність залишається досить низькою, крім моделі MLPClassifier (MLP) та RandomForestClassifier (RFC) (для даних з синтетичними зразками f1-score 0.85 та 0.74 відповідно при roc-auc 1 для обох моделей).

Під час аналізу моделей, навчених на наборах із застосуванням алгоритмів балансування та класового зважування (рис. 4), моделі, навчені на синтетичних даних, показали вищі показники точності. При цьому найгіршим алгоритмом боротьби з дисбалансом виявилося класове зважування, за якого максимальна точність f1-score не перевищувала 0.66 (DecisionTreeClassifier (DTC), датасет з синтетичними даними).

Хороші результати зафіксовані при застосуванні алгоритму випадкової недостатньої вибірки як для моделей, навчених на наборі з синтетичними зразками, так і без синтетичних зразків (за винятком MNB). Точність f1-score коливалася в межах 0.66-0.83 для даних лише з реальних зразків та 0.76-0.95 для даних з синтетичними зразками. Алгоритм випадкової надмірної вибірки, застосований на наборі з синтетичними прикладами, теж виявився корисним для побудови точних класифікаторів (f1-score знаходилося в діапазоні 0.74-0.88). Проте най-

Balance method	Model	Real data				Real+synthetic data			
		roc-auc	precision	recall	f1-score	roc-auc	precision	recall	f1-score
random oversampling	DTC	0,72	0,96	0,45	0,59	0,83	0,97	0,67	0,79
	KNC	0,64	0,96	0,20	0,33	0,86	0,97	0,65	0,78
	LR	0,95	0,97	0,57	0,71	0,97	0,98	0,83	0,90
	MLP	0,94	0,80	0,16	0,25	0,99	1,00	0,79	0,88
	MNB	0,95	0,80	0,18	0,29	0,98	1,00	0,60	0,74
	RFC	0,92	0,20	0,03	0,05	0,99	1,00	0,71	0,83
	SVC	0,90	0,81	0,77	0,77	0,91	0,92	0,65	0,76
SMOTE-NC	DTC	0,76	0,96	0,54	0,66	0,90	0,98	0,82	0,89
	KNC	0,85	0,94	0,65	0,76	0,92	0,96	0,80	0,87
	LR	0,97	0,98	0,48	0,63	0,99	0,98	0,89	0,93
	MLP	0,97	0,96	0,16	0,27	1,00	1,00	0,88	0,94
	MNB	0,96	1,00	0,21	0,33	0,99	1,00	0,65	0,78
	RFC	0,95	0,20	0,04	0,06	1,00	1,00	0,82	0,90
	SVC	0,93	0,81	0,86	0,82	0,92	0,92	0,65	0,76
random downsampling	DTC	0,69	0,67	0,80	0,72	0,82	0,82	0,82	0,82
	KNC	0,83	0,87	0,80	0,82	0,86	0,92	0,75	0,82
	LR	0,87	0,80	0,86	0,82	0,94	0,94	0,81	0,86
	MLP	0,78	0,85	0,66	0,66	0,99	0,80	0,73	0,76
	MNB	0,86	0,00	0,00	0,00	0,92	0,60	0,02	0,03
	RFC	0,88	0,71	1,00	0,82	0,99	0,93	0,97	0,95
	SVC	0,86	0,86	0,83	0,83	0,91	0,95	0,65	0,77
class weighting	DTC	0,68	0,22	0,37	0,27	0,82	0,66	0,66	0,66
	KNC	0,64	0,21	0,20	0,20	0,85	0,61	0,65	0,63
	LR	0,95	0,60	0,09	0,15	0,96	0,88	0,45	0,60
	MNB	0,89	0,00	0,00	0,00	0,90	0,50	0,02	0,03
	RFC	0,97	0,00	0,00	0,00	0,99	0,98	0,67	0,80
	SVC	0,93	0,05	0,80	0,09	0,91	0,38	0,66	0,48

Рис. 4. Теплова карта метрик точностей побудованих класифікаційних моделей

кращим методом балансування є алгоритм SMOTE-NC. Recall моделей, навчених на даних з синтетичними зразками, збалансованих за допомогою цього алгоритму, перевищував recall моделей, навчених на аналогічних вихідних даних, які були збалансовані випадковою надмірною вибіркою (0.65-0.89 проти 0.60-0.83 відповідно). Дві з 7 моделей (MLP, RFC) досягли максимального значення roc-auc, наближеного до 1 з високим значення precision (теж приблизно 1). Загальна точність варіювалася від 0.76 до 0.94. MLPClassifier є найточнішою моделлю (roc-auc 0.997) – має найвищі значення recall (0.9) та f1-score (0.94) серед усіх інших моделей та дуже високе значення precision (0.99). Це свідчить про те, що MLPClassifier має найкращу здатність розпізнавати позитивні класи, тобто випадки шахрайства. З матриці плутанини (рис. 5) видно, що кількість хибно-негативних результатів для позитивного класу не перевищує 10% від загальної кількості шахрайських оголошень.

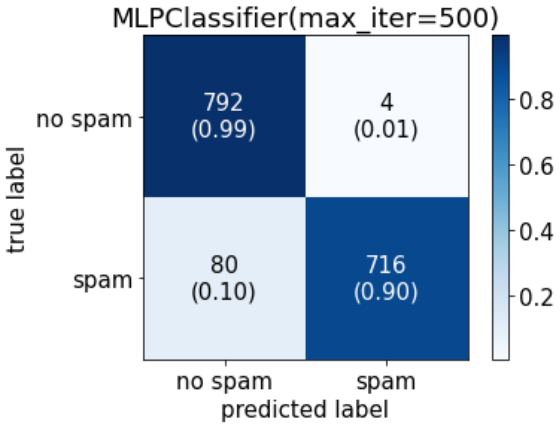

Рис. 5. Матриця плутанини для моделі MLPClassifier, навченій на збалансованих методом SMOTE-NC даних з синтетичними зразками

Було проведено аналіз хибно-негативних за природою походження даних, зважаючи на те, що для побудови класифікаторів були використані синтетичні дані.

З рис. 6 можна побачити, що найчастіше моделі помиляються при класифікації реальних шахрайських повідомлень та шахрайських повідомлень, згенерованих за допомогою ChatGPT. Дані, згенеровані нейронною мережею LSTM, класифікуються відносно добре. Це може бути пов'язано з тим, що реальні зразки та зразки, отримані з ChatGPT, більш

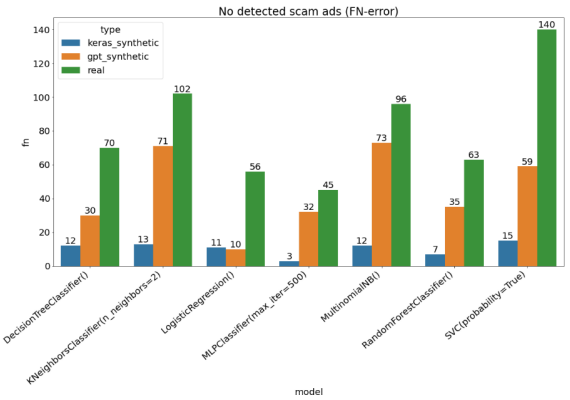

Рис. 6. Розподіл помилок другого роду для класифікаційних моделей за природою даних

різноманітні та складніші. У підсумку, найкращою комбінацією для розпізнавання шахрайських оголошень про оренду є алгоритм балансування SMOTE-NC, застосований на датасеті з синтетичними зразками, та модель MLPClassifier, навчена на даному датасеті. За допомогою такого поєднання можна створити ефективний метод виявлення і протидії фішингу.

4. Висновки

У ході цього дослідження для виявлення шахрайських оголошень про оренду було побудовано 70 класифікаційних моделей, на основі даних з синтетичними та без синтетичних зразків без застосувань алгоритмів балансування та із застосуванням випадкової надмірної вибірки, SMOTE-NC, випадкової недостатньої вибірки та класового зважування. За результатами аналізу точностей побудованих моделей було визначено, що використання синтетичних зразків значно поліпшує точність класифікації як на незбалансованих даних, так і для кожного алгоритму балансування, та для кожного класифікаційного алгоритму. Методи балансування, пов'язані зі зміною розміру датасету, виявилися ефективнішими за класове зважування. Найкращим алгоритмом балансування визначено метод SMOTE-NC, оскільки метрики точності моделей, навчених на наборі даних, збалансованому даним методом, були найвищими. Найточнішою та найефективнішою моделлю для розпізнавання шахрайських оголошень за результатами порівняльного аналізу метрик точності, виявилася модель MLPClassifier з roc-auc 0.997, recall 0.9, precision 0.99 та f1-score .94, що є найкращими показниками серед усіх досліджених класифікаторів. Таким чином розроблена модель може допомогти користувачам ефективно розпізнавати оголошення про оренду з ознаками шахрайства і зменшити ризик потенційної шахрайської діяльності.

Можливим напрямком подальших досліджень є використання запропонованого методу для розпізнавання шахрайських оголошень в інших сферах. Наприклад, в оголошеннях про роботу, про продаж товарів з різних категорій, про пересилання та кур'єрські послуги, тощо. Таким чином, дослідження може мати значний вплив на більшість індустрій та сприяти покращенню безпеки для користувачів.

Перелік використаних джерел

1. *ОПОРА*. Вплив повномасштабної війни на міграцію українців: як масштаби переміщення оцінюють держава Україна та міжнародні організації. — 15.02.2023.
2. *OLX.ua Б.* Шахраям – ні: що варто знати про пошук та оренду житла під час війни. — 18.04.2022.
3. *Oshikawa R., Qian J., Wang W. Y.* A survey on natural language processing for fake news detection // arXiv preprint arXiv:1811.00770. — 2018. — DOI: 10.48550/arXiv.1811.00770. — URL: https://doi.org/10.48550/arXiv.1811.00770.
4. *Jabbar M., Suharjito S.* Fraud Detection Call Detail Record Using Machine Learning in Telecommunications Company // Advances in Science, Technology and Engineering Systems Journal. — 2020. — Лип. — Т. 5. — С. 63—69. — DOI: 10.25046/aj050409.
5. A Natural Language Processing Approach for Financial Fraud Detection / J. F. Rodríguez, M. Papale, M. Carminati, S. Zanero [та ін.] // Proceedings of the Italian Conference on Cybersecurity ITASEC 2022, Rome, Italy, June 20-23, 2022. Т. 3260. — CEUR-WS. org. 2022. — С. 135—149.
6. *Omar S. J., Fred K., Swaib K. K.* A state-of-the-art review of machine learning techniques for fraud detection research // Proceedings of the 2018 International Conference on Software Engineering in Africa. — 2018. — С. 11—19.
7. Detection of fake job postings by utilizing machine learning and natural language processing approaches / A. Amaar, W. Aljedaani, F. Rustam, S. Ullah, V. Rupapara, S. Ludi // Neural Processing Letters. — 2022. — С. 1—29.
8. *Rani V.* Notification system to show spam free ads for housing rentals on Craigslist. — 02/06/2019.