УДК 004.85:004.738.5
DOI https://doi.org/10.24144/2616-7700.2026.48(1).137-145

В. В. Ковордaній

ДВНЗ «Ужгородський національний університет»,
аспірант кафедри системного аналізу та теорії оптимізації
volodymyr.kovordanii@uzhnu.edu.ua
ORCID: https://orcid.org/0009-0005-2097-4392

ГІБРИДНІ АРХІТЕКТУРИ ГЛИБОКОГО НАВЧАННЯ ДЛЯ КЛАСИФІКАЦІЇ ВЕБ-КОНТЕНТУ

У статті розглянуто проблему класифікації веб-контенту, що має критичне значення в умовах експоненційного зростання цифрових даних та є фундаментальною задачею обробки природної мови. Традиційні моделі глибокого навчання, попри свою ефективність, мають певні обмеження, що стимулювало розвиток гібридних архітектур. Метою даної статті є огляд гібридних архітектур глибокого навчання за останнє десятиліття.

Методологія дослідження включає огляд та порівняльний аналіз ключових підходів, починаючи від фундаментальних комбінацій згорткових (CNN) та рекурентних (RNN) нейронних мереж, через моделі, посилені механізмами уваги, до сучасних архітектур на основі трансформерів, графових нейронних мереж (GNN) та мультимодальних моделей, що інтегрують текст, DOM-структуру та візуальні ознаки. Показано еволюцію від текстових моделей до інтеграції потужних попередньо навчених мовних моделей (PLM), таких як BERT, що виступають у ролі основи для гібридних класифікаторів та графово-мультимодальних рішень.

Встановлено, що сучасні гібридні архітектури, особливо ті, що використовують трансформери та враховують структурні й мультимодальні аспекти веб-контенту, демонструють найвищу ефективність, проте водночас ставлять нові виклики, пов'язані з обчислювальною складністю, інтерпретованістю та дефіцитом сучасних відкритих датасетів саме для веб-класифікації.

Ключові слова: класифікація веб-контенту, глибоке навчання, гібридні архітектури, обробка природної мови, згорткові нейронні мережі (CNN), рекурентні нейронні мережі (RNN), трансформери, графові нейронні мережі (GNN), мультимодальність.

1. Вступ. Класифікація веб-контенту є однією з ключових задач сучасного інформаційного суспільства. На відміну від класичного тексту, веб-сторінка є напівструктурованим мультимодальним об'єктом, який окрім тексту містить зображення, гіперпосилання та ієрархічну структуру, визначену DOM-деревом.

Ранні методи машинного навчання вимагали ручного створення ознак, тоді як глибоке навчання автоматизувало цей процес, значно підвищивши точність [1, 2]. Однак окремі архітектури глибокого навчання мають свої обмеження. Згорткові нейронні мережі (CNN), запозичені з галузі комп'ютерного зору, ефективно виділяють локальні ознаки, такі як n-грами, але є інваріантними до порядку слів і не здатні вловлювати довгострокові залежності в тексті [3]. З іншого боку, рекурентні нейронні мережі (RNN) та їхні вдосконалені варіанти, як-от Long Short-Term Memory (LSTM) та Gated Recurrent Unit (GRU), добре моделюють послідовності, але можуть втрачати інформацію з віддалених частин тексту через проблему зникаючого градієнта та менш ефективно виділяють локальні патерни [4]. Це стало рушійною силою для створення гібридних архітектур, які поєднують переваги різних моделей.

Під гібридною архітектурою розуміємо модель, що поєднує різні класи нейромережевих підходів або джерела ознак (наприклад, текст + зображення, текст + структура DOM) для компенсації обмежень окремих компонентів. Під мультимодальністю розуміємо спільне використання кількох модальностей вебсторінки (тексту, DOM структури, візуальних ознак тощо).

Метою роботи є проведення комплексного огляду найвпливовіших гібридних підходів за останнє десятиліття, аналіз їхніх архітектурних рішень та визначення перспективних напрямків досліджень, зокрема в галузі графових та мультимодальних систем.

Стаття має оглядовий характер. Наведені метрики узагальнюють результати публікацій і не є результатом уніфікованого експерименту на одному датасеті.

2. Аналіз останніх досліджень і публікацій. Еволюцію гібридних архітектур можна умовно поділити на декілька етапів, що характеризуються впровадженням нових технологій. Варто відзначити, що не всі розглянуті методи враховують мультимодальність веб-контенту, а зосереджені виключно на проблемі класифікації текстової складової.

2.1. Фундаментальні гібридні моделі. Перші гібридні архітектури поєднували згорткові (CNN) та рекурентні (RNN) мережі для синергетичного ефекту: CNN виділяли локальні ознаки (n-грами), а RNN моделювали довгострокові залежності в тексті. Піонерські роботи, такі як C-LSTM, продемонстрували, що такий підхід, де вихід CNN подається на вхід LSTM, дозволяє одночасно враховувати як локальні, так і глобальні семантичні патерни [5]. Важливо, що показники C-LSTM-подібних CNN+RNN моделей суттєво залежать від гіперпараметрів: розміру ядра CNN і кількості фільтрів (що визначає "локальність" ознак), довжини усічення послідовності, розміру прихованого стану LSTM/GRU, dropout та стратегії пулінгу. Саме різні налаштування і правила препроцесингу часто пояснюють розбіжності у результатах.

Альтернативний підхід був представлений у моделі R-CNN [6]. Тут рекурентна структура (зазвичай двонаправлена RNN) використовується для збагачення векторного представлення кожного слова інформацією про його лівий та правий контекст. Лише після цього до збагачених векторів застосовується згортковий шар та операція пулінгу для отримання фінального вектора ознак. Це дозволяє згортковим фільтрам працювати не з ізольованими словами, а з контекстуалізованими представленнями. Для R-CNN приріст якості зазвичай з'являється тоді, коли контекст навколо маркерів класу важливіший за сам факт їх присутності. Водночас обчислювально рекурентна частина ускладнює паралелізацію, тому за однакової точності прості CNN можуть бути практично вигіднішими для швидкого інференсу.

Наступним кроком стало впровадження механізмів уваги, які дозволили моделям динамічно зважувати важливість різних частин тексту. Яскравим прикладом є ієрархічна мережа уваги (HAN), що застосовує увагу на рівнях слів та речень, що не тільки підвищило точність, але й забезпечило кращу інтерпретованість рішень [7].

2.2. Гібриди на основі трансформерів. Поява великих попередньо навчених мовних моделей (PLM), зокрема BERT, спричинила революцію в NLP. Домінуючою стала парадигма трансферного навчання, коли модель, попередньо навчена на величезних масивах нерозмічених текстових даних (наприклад,

уся Вікіпедія), доналаштовується під конкретну прикладну задачу на відносно невеликому наборі розмічених даних [8].

BERT та його аналоги використовуються як потужні енкодери, що генерують глибоко контекстуалізовані векторні представлення слів. Це призвело до появи нового класу гібридних архітектур, де BERT виступає в ролі базового енкодера, а "класичні"архітектури, такі як CNN та RNN, використовуються як ефективні "класифікаційні голови"для обробки збагачених контекстом ембедингів. Найбільш поширеними є комбінації BERT-CNN [9], BERT-LSTM/GRU та складніші архітектури, як-от BERT-BiLSTM-CNN [10, 11]. Такі підходи дозволяють одночасно враховувати контекст, послідовність та локальні патерни, що часто призводить до найвищої точності на складних задачах класифікації.

Для BERT-орієнтованих гібридів покращення метрик часто пов'язане з fine-tuning попередньо навчених моделей, що підвищує вимоги до GPU-ресурсів і часу навчання.

2.3. Гібридні моделі на основі графових нейронних мереж (GNN). Новим перспективним напрямком є використання графових нейронних мереж (GNN) [12], які представляють текстові дані у вигляді графової структури, де вузлами є слова або документи, а ребра відображають їхні взаємозв'язки. Моделі, такі як TextGCN, перетворюють задачу класифікації тексту на задачу класифікації вузлів у гетерогенному графі, що дозволяє враховувати глобальну структуру всього корпусу даних [13]. Гібридизація GNN з трансформерами (наприклад, BERTGCN) виявилася особливо ефективною, поєднуючи потужність PLM для отримання глибоких ембедингів зі здатністю GNN моделювати складні структурні зв'язки [14]. Але в той же час є однією з найбільш ресурсомістких архітектур, що вимагає велике споживання відеопам'яті для BERT та високі вимоги до оперативної пам'яті для графової структури.

2.4. Мультимодальні гібридні архітектури. Веб-контент за своєю природою є мультимодальним, тобто часто містить не лише текст, а й зображення, відео та аудіо. Мультимодальні гібридні архітектури спрямовані на вирішення цієї проблеми шляхом спільної обробки та інтеграції інформації з різних джерел. Основний принцип таких моделей полягає у використанні спеціалізованих енкодерів для кожної модальності (наприклад, CNN для зображень та BERT для тексту), а потім у злитті (fusion) отриманих представлень для фінальної класифікації.

Прикладом такої архітектури є модель HTIC (Hybrid Text Image Classifier), яка використовує VGG16 та оптимізовану CNN для класифікації зображень та RoBERTa для класифікації тексту, поєднуючи їхні виходи для отримання кінцевого результату [15].

Мультимодальні моделі характеризуються високою затримкою виводу. Процес виводу включає не лише прогін через нейронну мережу, але й попередню обробку різних модальностей: рендеринг веб-сторінки, OCR (оптичне розпізнавання символів) для виділення тексту з зображень та генерацію візуальних ознак. Довгий час обробки однієї сторінки робить такі моделі придатними переважно для офлайн-аналізу.

2.5. Графово-мультимодальні та структурно-орієнтовані рішення для веб-сторінок. Останній етап еволюції моделей для класифікації веб-контенту характеризується відходом від представлення веб-сторінок як про-

стого тексту та переходом до архітектур, що цілісно обробляють їхню складну природу, враховуючи HTML-структуру (DOM-дерево), візуальне розташування елементів та графові зв'язки.

Моделі, що враховують структуру, такі як DOM-LM [16] та MarkupLM [17], були одними з перших, хто спробував інтегрувати структурну інформацію в архітектуру Transformer. DOM-LM кодує локальну структуру DOM-дерева за допомогою спеціальних позиційних вкладень, що описують глибину вузла, індекс батьківського та сусідніх елементів. MarkupLM, у свою чергу, використовує вирази XPath для представлення глобального розташування кожного текстового токена в ієрархії документа.

Наступним кроком стала інтеграція візуальної модальності. LayoutLMv2 [18] є мультимодальною моделлю, що одночасно обробляє текст, інформацію про розташування (2D-координати) та зображення сторінки, що робить її ефективною для документів з фіксованою структурою, як-от PDF-файли. Для динамічних веб-сторінок була розроблена модель WebLM, яка вдосконалює цей підхід, використовуючи HTML-структуру для ієрархічного агрегування візуальних ознак. Це дозволяє моделі бути стійкою до адаптивних дизайнів, де абсолютні координати елементів можуть змінюватися [19].

Окремий напрямок представляють гібридні моделі PLM-GNN, які поєднують попередньо навчені мовні моделі (PLM) з графовими нейронними мережами (GNN). У такій архітектурі PLM (наприклад, BERT) відповідає за кодування текстового вмісту, а GNN безпосередньо моделює DOM-дерево як граф, що дозволяє ефективно враховувати його природну ієрархічну структуру [20].

Обмеженням структурно-орієнтованих підходів є залежність від якості HTML розмітки та стабільності верстки. Реальні веб-сторінки часто містять динамічні блоки, неконсистентну розмітку або шаблонні навігаційні елементи. Також зростає обчислювальна складність (парсинг DOM, побудова графа або витяг layout-ознак), що впливає на час інференсу й складність розгортання у реальному середовищі.

3. Формалізація задачі. З математичної точки зору, задача класифікації веб-контенту полягає у навчанні функції-класифікатора f, яка відображає веб-сторінку d з простору документів D в одну з категорій c із попередньо визначеної множини категорій $C = \{c_1, c_2, \ldots, c_m\}$. Формально це можна записати як:

$$f : D \to C$$

У контексті глибокого навчання цей процес реалізується через кілька послідовних етапів: векторизація (embedding), кодування (encoding) та класифікація (classification).

На першому етапі вхідний текстовий документ d перетворюється на послідовність числових векторів $X = (x_1, x_2, \ldots, x_k)$, де x_i — це векторне представлення (ембединг) i-го токена (слова або частини слова) в документі. Цей крок виконується за допомогою шару вкладень, який може бути навчений з нуля, ініціалізований попередньо навченими векторами (наприклад, Word2Vec, GloVe), або, що є найбільш сучасним підходом, отриманий з попередньо навченої мовної моделі, такої як BERT.

Послідовність векторів X подається на вхід гібридної архітектури глибокого навчання $g(X)$. Ця архітектура обробляє послідовність і перетворює її у

векторне представлення документа фіксованої довжини v, яке містить у собі всю релевантну для класифікації семантичну інформацію:

$$v = g(X)$$

Отриманий вектор документа v подається на повнозв'язний шар з функцією активації Softmax. Ця функція обчислює розподіл ймовірностей приналежності документа до кожної з m категорій. Ймовірність для j-ї категорії обчислюється за формулою:

$$P(y = j \mid d) = softmax(Wv + b)_j = \frac{e^{(w_j v + b_j)}}{\sum_{i=1}^{m} e^{(w_i v + b_i)}},$$

де W та b — це матриця вагових коефіцієнтів та вектор зміщення класифікаційного шару, які навчаються в процесі тренування моделі.

Навчання моделі здійснюється шляхом мінімізації функції втрат, якою зазвичай виступає перехресна ентропія (cross-entropy loss), між прогнозованим розподілом ймовірностей та істинною (one-hot encoded) міткою категорії.

Для графових моделей задача дещо видозмінюється. Замість окремого документа d, вхідними даними є цілий корпус, представлений у вигляді графа $G = (V, E)$, де вузли V — це і документи, і слова, а ребра E відображають їхні взаємозв'язки. Задача класифікації перетворюється на задачу класифікації вузлів (node classification) у цьому графі.

У випадку мультимодальної класифікації, вхідний об'єкт d складається з кількох компонентів, наприклад, тексту d_{text} та зображення d_{image}. Процес кодування включає окремі енкодери для кожної модальності, $g_{text}(X)$ та $g_{image}(X)$, а фінальний вектор v отримується шляхом злиття (fusion) їхніх виходів:

$$v = fuse(g_{text}(X_{text}), g_{image}(X_{image}))$$

4. Порівняльна характеристика архітектур. Важливим аспектом оцінки ефективності гібридних архітектур є аналіз експериментальної бази. Результати часто наводяться на різнорідних наборах даних, які можна розділити на дві фундаментально відмінні групи: класичні текстові корпуси (Stanford Sentiment Treebank, R8, Yahoo Answers, SemEval-2019) та спеціалізовані веборієнтовані датасети (SWDE, WebSRC, SROIE, NFT). Розуміння відмінностей між ними є критичним для коректної інтерпретації показників точності.

Перша група наборів даних фокусується виключно на семантичному аналізі текстового вмісту, ігноруючи візуальну та структурну специфіку веб-сторінок.

Друга група датасетів розроблена спеціально для оцінки здатності моделей обробляти реальну природу веб-контенту, включаючи структуру DOM та візуальне розташування елементів.

Наведені у порівняльній таблиці 1 числові показники точності (Accuracy та F1-score) запозичені з різних джерел і отримані на різних датасетах, тому вони слугують ілюстрацією потенціалу методів, але не є прямим міжмодельним порівнянням.

5. Висновки та перспективи подальших досліджень. Проведений огляд демонструє еволюцію підходів до класифікації веб-контенту від текстоорієнтованих моделей до архітектур, що системно структуру DOM та візуальне

Таблиця 1.

Порівняльний аналіз архітектур

Архітектура	**Ключові особливості**	**Використані модальності**	**Результат на датасеті (Acc / F1)**	**Складність (Ресурси)**
CNN+RNN (C-LSTM)	Послідовне застосування: CNN для локальних ознак (n-грам) та LSTM для довгострокових залежностей.	Текст	**Stanford Sentiment Treebank** (двокласова класифікація): ∼87.8% (Acc) [5]	Середня
HAN	Ієрархічна увага на рівні слів та речень для кращої інтерпретованості та точності.	Текст (ієрархія)	**Yahoo Answers:** 75.8% (Acc) [7]	Середня
BERT-CNN	BERT як енкодер, CNN як класифікаційна голова для виділення локальних патернів.	Текст	**SemEval-2019 Task3:** 94.7% (Acc), 94% (F1) [9]	Висока
TextGCN	Моделювання всього корпусу як гетерогенного графу "слово-документ". Класифікація тексту як задача класифікації вузлів.	Текст + Граф (слова/документи)	**R8 (Reuters):** ∼97.07% (Acc) [13]	Висока (Пам'ять)
BertGCN	Граф слів BERT (для ембедингів вузлів-документів) та GCN (для поширення міток по графу).	Текст + Граф (слова/документи)	**R8 (Reuters):** 98.1% (Acc) [14]	Дуже Висока
HTIC	Мультимодальна модель з окремими енкодерами для тексту (RoBERTa) та зображень (VGG16, CNN).	Текст + Зображення	**NFT dataset:** >98% (Acc) [15]	Екстремальна
DOM-LM	PLM, що кодує локальну структуру DOM-дерева через позиційні ембедінги (глибина, індекс батька/сусіда).	Текст + Структура (DOM-дерево)	**SWDE:** ∼94.2% (F1) [16]	Висока
LayoutLMv2	Мультимодальна PLM, що одночасно обробляє текст, 2D-координати та візуальне зображення документа (через CNN).	Текст + Візуал (зображення/макет)	**SROIE:** 97.81% (F1) [18]	Висока
WebLM	Мультимодальна PLM для веб-сторінок, що використовує HTML-структуру для ієрархічної агрегації візуальних ознак.	Текст + Структура (DOM) + Візуал (зображення)	**WebSRC:** 78.66% (Acc) [19]	Висока
PLM+GNN	Спільне кодування тексту (PLM) та структури DOM-дерева (GNN).	Текст + Граф (DOM-дерево)	**SWDE:** 90.2% (Acc), 89.7% (F1) [20]	Висока

представлення сторінки. На основі наведених у таблиці прикладів можна сформулювати такі висновки (з урахуванням того, що метрики отримано на різних датасетах):

- Текстові гібриди є відносно легкими для впровадження та можуть бути практичним компромісом між якістю і складністю. Водночас їх результати суттєво залежать від гіперпараметрів і довжини тексту: CNN виграють на локальних сигналах, LSTM — на довгих залежностях, а C-LSTM — на задачах, де потрібні обидва типи ознак.
- Гібриди на основі трансформерів часто демонструють вищу точність, але

вимагають суттєво більших обчислювальних ресурсів та акуратної адаптації до домену веб-даних. Тому їх практична доцільність найбільша у сценаріях, де точність критична і допустимі витрати на ресурси.

- Графові та структурні підходи підкреслюють важливість глобальних зв'язків і структури документа. Вони перспективні саме для "веб-специфічних" задач, але потребують інженерно складнішого пайплайну (парсинг DOM, побудова графів, стійкість до помилок розмітки) та спеціалізованих датасетів.
- Мультимодальні гібриди мають великий потенціал, але їх перевага залежить від доступності якісних мультимодальних датасетів (текст + DOM + скріншоти) та ресурсів на навчання. Високі показники, подібні до наведених для HTIC, слід трактувати як результат у специфічних умовах експерименту, а не як універсальну перевагу для будь-якої веб-класифікації. Водночас через значну ресурсомісткість мультимодальні гібриди є неефективними для задач, що вимагають обробки потоків даних у режимі реального часу.

Підвищення якості в гібридних та мультимодальних системах зазвичай супроводжується збільшенням обчислювальної складності й ресурсозатратності. Майбутні дослідження мають бути спрямовані на розробку методів дистиляції знань, квантизації та розрідження для створення менших, але ефективних моделей, що здатні працювати у режимі реального часу.

Існує гостра потреба у створенні відкритих, багатомовних (зокрема україномовних) корпусів з еталонними розмітками класів та доступом до HTML, CSS та скріншотів для адекватного тестування мультимодальних моделей. Оскільки класичні датасети не репрезентують сучасний веб.

Також критично важливою є систематична оцінка надійності архітектур в умовах реального, "зашумленого"вебу. Моделі повинні бути стійкими до помилок у HTML-розмітці та інших артефактів, що є запорукою їх успішного практичного застосування.

Конфлікт інтересів

Автор заявляє, що не мають конфлікту інтересів щодо даного дослідження, включаючи фінансовий, особистий, авторський або будь-який інший, який міг би вплинути на дослідження, а також на результати, представлені в даній статті.

Фінансування

Дослідження було проведено без фінансової підтримки.

Доступність даних

Усі дані доступні в цифровій або графічній формі в основному тексті рукопису.

Використання штучного інтелекту

Автор підтверджує, що при створенні даної роботи він не використовував технології штучного інтелекту.

Список використаної літератури

1. Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In C. C. Aggarwal & C. Zhai (Eds.), Mining text data (pp. 163-222). Springer. Retrieved from https://scispace.com/pdf/a-survey-of-text-classification-algorithms-29nuhpcf91.pdf
2. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. Retrieved from https://www.cs.toronto.edu/ hinton/absps/NatureDeepReview.pdf
3. Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Retrieved from https://doi.org/10.48550/arXiv.1408.5882
4. Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. Retrieved from https://doi.org/10.48550/arXiv.1605.05101
5. Zhou, C., Sun, C., Liu, Z., & Lau, F. C. (2015). A C-LSTM Neural Network for Text Classification. ArXiv. Retrieved from https://doi.org/10.48550/arXiv.1511.08630
6. Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9745
7. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1480–1489). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/N16-1174
8. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv. Retrieved from https://doi.org/10.48550/arXiv.1810.04805
9. Abas, A. R., Elhenawy, I., Zidan, M., & Othman, M. (2022). BERT-CNN: A deep learning model for detecting emotions from text. Computers, Materials & Continua, 71(2), 2943–2961. Retrieved from https://doi.org/10.32604/cmc.2022.021671
10. Gou, Z., & Li, Y. (2023). Integrating BERT embeddings and BiLSTM for emotion analysis of dialogue. Computational Intelligence and Neuroscience, 2023, 6618452. Retrieved from https://doi.org/10.1155/2023/6618452
11. Xiong, Y., Chen, G., & Cao, J. (2024). Research on public service request text classification based on BERT-BiLSTM-CNN feature fusion. Applied Sciences, 14(14), 6282. Retrieved from https://doi.org/10.3390/app14146282
12. Wang, K., Ding, Y., & Han, S. C. (2023). Graph Neural Networks for Text Classification: A Survey. ArXiv. Retrieved from https://doi.org/10.1007/s10462-024-10808-0
13. Yao, L., Mao, C., & Luo, Y. (2018). Graph Convolutional Networks for Text Classification. ArXiv. Retrieved from https://doi.org/10.48550/arXiv.1809.05679
14. Lin, Y., Meng, Y., Sun, X., Han, Q., Kuang, K., Li, J., & Wu, F. (2021). BertGCN: Transductive Text Classification by Combining GCN and BERT. ArXiv. Retrieved from https://doi.org/10.48550/arXiv.2105.05727
15. Gupta, S., & Kishan, B. (2025). A performance-driven hybrid text-image classification model for multimodal data. Scientific Reports, 15, 11598. Retrieved from https://doi.org/10.1038/s41598-025-95674-8

16. Deng, X., Shiralkar, P., Lockard, C., Huang, B., & Sun, H. (2022). DOM-LM: Learning Generalizable Representations for HTML Documents. ArXiv. Retrieved from https://doi.org/10.48550/arXiv.2201.10608
17. Li, J., Xu, Y., Cui, L., & Wei, F. (2021). MarkupLM: Pre-training of Text and Markup Language for Visually-rich Document Understanding. ArXiv. Retrieved from https://doi.org/10.48550/arXiv.2110.08518
18. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., Zhang, M., & Zhou, L. (2020). LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding. ArXiv. Retrieved from https://doi.org/10.48550/arXiv.2012.14740
19. Xu, H., Chen, L., Zhao, Z., Ma, D., Cao, R., Zhu, Z., & Yu, K. (2024). Hierarchical Multimodal Pre-training for Visually Rich Webpage Understanding. ArXiv. Retrieved from https://doi.org/10.48550/arXiv.2402.18262
20. Lang, Q., Zhou, J., Wang, H., Lyu, S., & Zhang, R. (2023). PLM-GNN: A Webpage Classification Method based on Joint Pre-trained Language Model and Graph Neural Network. ArXiv. Retrieved from https://doi.org/10.48550/arXiv.2305.05378

Kovordaniy V. V. Hybrid deep learning architectures for web content classification.

This paper addresses the problem of web content classification, a task of critical importance amidst the exponential growth of digital data and a fundamental challenge in Natural Language Processing (NLP). Despite their effectiveness, traditional deep learning models possess certain limitations, which has necessitated the development of hybrid architectures. The aim of this paper is to review hybrid deep learning architectures developed over the past decade.

The research methodology involves a review and comparative analysis of key approaches, ranging from fundamental combinations of Convolutional (CNN) and Recurrent (RNN) Neural Networks, through models enhanced by attention mechanisms, to state-of-the-art architectures based on Transformers, Graph Neural Networks (GNNs), and multimodal models integrating text, DOM structure, and visual features. The study demonstrates the evolution from text-only models to the integration of powerful Pre-trained Language Models (PLMs), such as BERT, which serve as the backbone for hybrid classifiers and graph-multimodal solutions.

It is established that modern hybrid architectures, particularly those utilizing Transformers and incorporating structural and multimodal aspects of web content, exhibit superior performance. However, they concurrently present new challenges regarding computational complexity, interpretability, and the scarcity of up-to-date open datasets specifically designed for web classification.

Keywords: web content classification, Deep Learning, hybrid architectures, Natural Language Processing, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Transformers, Graph Neural Networks (GNN), multimodality.

Отримано: 03.10.2025 Прийнято: 12.11.2025 Опубліковано: 29.01.2026