УДК 004.934:316.77

АНАЛІЗ МЕТОДОЛОГІЧНИХ ПІДХОДІВ ПРОГНОЗУВАННЯ ВІРУСНОСТІ КОНТЕНТУ ЗА ДОПОМОГОЮ BIG DATA

Калініна О.І., Сиромятникова Д.О., Супрун О.О.
e-mail: oleksandra.kalinina@nure.ua, daria.syromiatnykova@nure.ua
Харківський національний університет радіоелектроніки, каф. МСТ
м. Харків, Україна

A comprehensive analysis of scientific research on predicting the virality of online content using big data analysis methods has been conducted in the article. The evolution of methodological approaches is examined: from classical time series-based models to modern deep learning architectures, including graph neural networks, transformers, and multimodal models. Special attention is paid to comparative analysis of methods, problems of early prediction, cross-platform content dissemination, model interpretability, and overcoming prediction drift.

Феномен вірусного поширення контенту в соціальних медіа став невід'ємною характеристикою сучасного інформаційного простору. Здатність прогнозувати, який саме контент набуде широкого розповсюдження, має критичне значення для цифрового маркетингу, медіабізнесу та управління репутацією брендів. Водночас стрімке зростання обсягів даних, генерованих користувачами соціальних платформ, створює як безпрецедентні можливості для аналітики, так і серйозні виклики для дослідників.

Метою цієї статті є систематизація та критичний аналіз сучасних наукових досліджень, присвячених прогнозуванню вірусності контенту за допомогою методів аналізу великих даних, з фокусом на порівняльному аналізі методологічних підходів та їхньої ефективності в різних умовах застосування.

Початкові дослідження прогнозування популярності контенту базувалися на класичних статистичних методах (табл. 1). Проте ці підходи мали суттєві обмеження: залежність від ручного інжинірингу ознак, нереалістичні припущення про незалежність спостережень та недостатня ефективність для роботи зі складними каскадами поширення.

Як показали подальші дослідження, точність таких моделей суттєво знижувалася при перенесенні на нові платформи або типи контенту.

Наступний етап розвитку дослідницького поля пов'язаний із моделюванням каскадів поширення структур, що утворюються в результаті репостів та шерингу контенту.

Використовуючи дані з Twitter, дослідники довели, що за наявності 50% даних про вірусність контенту похибка прогнозованого часу не перевищує 40%.

Таблиця 1 – Порівняння класичних статистичних методів прогнозування вірусності

Метод	Переваги	Обмеження	Типове застосування
1	2	3	4
Логарифмічно-нормальний розподіл	Простота реалізації, інтерпретованість	Нереалістичні припущення про незалежність спостережень	Довгострокове прогнозування популярності відео
Лінійна/ логістична регресія	Прозорість моделі, низька обчислювальна вартість	Обмежена здатність моделювати нелінійні залежності	Прогнозування популярності новин
Методи на основі каскадів	Врахування соціальної динаміки	Складність отримання повної структури каскаду	Аналіз поширення в Twitter

Сучасний медіаконтент характеризується мультимодальною природою: текст поєднується з зображеннями, відео, аудіо та інтерактивними елементами. Порівняльний аналіз демонструє, що інтеграція декількох модальностей суттєво підвищує точність прогнозування порівняно з унімодальними підходами (табл. 2).

Таблиця 2 – Порівняння унімодальних та мультимодальних підходів до прогнозування вірусності

Набір модальностей	Точність (Accuracy)	Повнота (Recall)	F1-score	Пояснювальна здатність
1	2	3	4	5
Тільки текст	0.72	0.68	0.70	Висока (семантичний аналіз)
Тільки зображення	0.65	0.62	0.63	Середня (візуальні патерни)
Тільки часові ознаки	0.70	0.75	0.72	Низька (чорний ящик)
Текст + зображення	0.81	0.78	0.79	Середня
Текст + часові ознаки	0.83	0.80	0.81	Середня
Повна мультимодальність	0.89	0.87	0.88	Низька

Висновки. Прогнозування вірусності контенту за допомогою методів аналізу великих даних є динамічною та міждисциплінарною галуззю досліджень, що швидко розвивається. За останнє десятиліття відбулася еволюція від простих регресійних моделей до складних архітектур глибинного навчання, здатних враховувати структурні, часові та мультимодальні характеристики контенту та його поширення.

Проведений порівняльний аналіз демонструє, що жоден метод не є універсальним оптимумом. Мультимодальні підходи забезпечують комплексне розуміння контенту, але створюють додаткові виклики у вирівнюванні різних модальностей. Залишаються відкритими важливі виклики: потреба в інтерпретованих моделях, проблема гетерогенності платформ, складність отримання якісних даних та необхідність адаптації моделей до різних типів контенту. Подальші дослідження мають бути спрямовані на розробку гібридних підходів, що поєднують переваги різних методів, та створення уніфікованих фреймворків, здатних працювати в реальному часі в умовах сучасного гібридного медіасередовища.

Список використаних джерел:

1. Szabo, G., & Huberman, B.A. (2010). Predicting the popularity of online content. Communications of the ACM, 53(8), 80-88. DOI: https://doi.org/10.1145/1787234.1787238.

2. Kong, Q., Rizoiu, M.A., & Xie, L. (2020). Describing and predicting online items with reshare cascades via dual mixture self-exciting processes. Proceedings of the 29th ACM International Conference on Information and Knowledge Management. New York. (p. 645-654). DOI: https://doi.org/10.1145/3340531.3411885.

3. Драннік, А. (2025). Застосування генеративних моделей AI для обробки медіа в реальному часі. Автоматизація та приладобудування. (с. 127-131).

4. Горбачов, К. (2025). Інтеграція штучного інтелекту в медіаіндустрію. Автоматизація та приладобудування. (с. 121-126).

5. Супрун, О.О., & Чалик, Д.С. (2025) ШІ для аналізу великих обсягів даних з соціальних мереж та новинних сайтів для виявлення трендів. Поліграфічні, мультимедійні та web-технології. Т. 2. (с. 14-16)