

УДК 004.8

Ховрат А.В.¹, Кобзев В.Г.²

¹ асп. Харківського національного університету радіоелектроніки

² проф. Харківського національного університету радіоелектроніки

ДВОШАРОВА МОДЕЛЬ КЛАСИФІКАЦІЙ СФАБРИКОВАНИХ ДАНИХ В СОЦІАЛЬНО ОРІЄНТОВАНИХ СИСТЕМАХ

Проблема виявлення сфабрикованої інформації в системах, де між собою взаємодіють багато користувачів, наприклад, у соціальних мережах, усе частіше розглядається як в академічній, так і у правовій сферах [1]. Це пов'язано з розвитком технологій генерації та загальним збільшенням інформаційного навантаження на індивіда. У критичних ситуаціях при соціальних зсувах, дані подібного роду здатні суттєво зашкодити окремо кожній людині, так і соціуму загалом. З метою уникнення суб'єктивної оцінки сприйняття інформації, є доцільним розглядати залучення засобів інтелектуального аналізу даних. Особливості реалізації тих чи інших підходів прямо залежать від характеру даних, які розглядаються. У рамках чинної роботи було вирішено обмежитися лише текстовими новинами. Це пов'язано з тим, що на момент проведення дослідження, відсутні інструменти створення достовірних, з точки зору людського ока, відео підробок. У той же час аудіо інформація мало поширена в соціальних мережах.

При класифікації текстових даних найбільш популярними є три ключових підходи [2]:

- використання ймовірнісних моделей, наприклад, наївний Байєсів класифікатор, ланцюги Маркова, Баєсові мережі тощо;
- використання нейронних мереж, наприклад, рекурентні чи згорткові мережі, трансформери чи інші глибинні моделі штучного інтелекту;

- використання наївних поліноміальних моделей, наприклад, із залученням лінійної адитивної згортки з ваговими коефіцієнтами та певним обмежувальними значенням.

У попередніх роботах, присвячених бінарній класифікації даних на достовірні та сфабриковані вже розглядалися ймовірнісні моделі та алгоритми із залученням різноманітних нейронних мереж [3, 4]. Результати показали, що однією із найбільш ефективних з точки зору точності та швидкодії є гібридна мережа, яка поєднує у собі рекурентні та згорткові можливості – RCNN. Окрім цього було виявлено, що однією з проблем при дослідженні подібного розмежування інформації є визначення ступеня важливості фейку. Зокрема, деякі тексти можуть носити відкрито гумористичний характер, який прямо сприймається людиною. Відповідно, суспільна шкода цих даних відсутня. У той же час новини спрямовані на дискредитацію тих чи інших важливих для соціуму законодавчих ініціатив – мають суттєвий рівень небезпеки.

Враховуючи зазначене вище, у рамках чинної роботи було вирішено сформувати набір класів фейкової інформації, які слугуватимуть базисом для проведення класифікації відсіяних за допомогою RCNN даних. Таким чином буде сформовано двошарова модель виявлення сфабрикованих новин з метою визначення найбільш ефективного підходу до розмежування подібних даних. Задля виконання поставленої мети було виокремлено перелік завдань:

- проведення експертного оцінювання та аналізу предметної галузі для визначення базових класів фейкової інформації;
- формування моделі моделей розбиття даних на групи за допомогою залучення наївного Байєсового класифікатора;
- проведення експериментальної перевірки запропонованої двошарової моделі у порівнянні з RCNN підходом.

Першим кроком вирішення завдання мультикласифікації є визначення основних класів дезінформації. Для визначення списку було вирішено залучити експертну групу з 100 аналітиків даних з різних країн Європи та Північної Америки. У результаті був складений список з 10 найпоширеніших категорій. Після цього було проведено відкрите опитування щодо виявлення найбільш вразливих видів фальсифікацій інформації. Зведені відповіді 300 учасників були використані для формування цих визначених груп:

- жарти з об'єктивно визначеними проявами;
- жарти з контекстуальними чи граматичними проявами;
- новини, орієнтовані на окремих осіб або невеликі групи;
- новини, орієнтовані на кілька регіонів, країн або великих груп;
- новини, орієнтовані на кілька країн або суспільство загалом.

Визначивши ознаки фейкової інформації, можна переходити до розробки набору характеристик, які будуть служити вхідними значеннями

для моделей. Першим таким показником є «Емоційна характеристика», яка виводиться на основі принципів контент-аналізу [5].

Крім цієї характеристики було вирішено розглянути ще 3 додаткові:

- індекс підозрілості: визначається за допомогою обчислення нормалізованої кількості слів із пов'язаного фейкового списку;

- вага повідомлення: дані будуть розділені на чотири групи: фрагменти домашнього діалогу, загальні новини, інформація про надзвичайні події та новини особливої важливості.

- індикатор дезінформації: визначається за допомогою 100 найпопулярніших статей новин, пов'язаних із вибраною темою, із застосуванням розпаралеленої версії алгоритму Apriori.

Після розгляду ключових характеристик можемо перейти до обговорення моделей, які використовуються для класифікації – RCNN. Ця модель поєднує кілька згорткових нейронних мереж, здатних працювати паралельно, з двонаправленою рекурентною мережею, яка підтримує як короткочасну, так і довготривалу пам'ять. У підході використовуються гіперболічні тангенси та сигмоїди, ефективно вирішуючи проблеми зникаючого та вибухового градієнтів шляхом обмеження діапазону результатуючих значень. Крім того, визначена модель дозволяє врахувати історичний контекст у повній мірі. Графічно її можна подати за допомогою схеми наведеної на Рисунку 1.

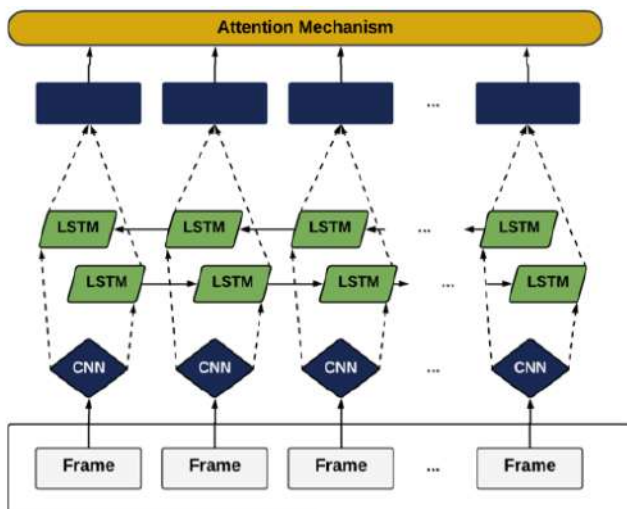


Рисунок 1 – Схематичне зображення архітектури RCNN.

Суть наївного Байєсового класифікатора полягає в застосуванні теореми Байєса для встановлення ймовірності належності спостереження до певного класу, припускаючи незалежність між змінними. Для підвищення ефективності даного підходу було вирішено розширити кількість показників на основі попередніх досліджень:

- співвідношення риторичних конструкцій до кількості речень;
- співвідношення негативних конструкцій до кількості речень;
- співвідношення вставних конструкцій до кількості речень.

Перейдемо до визначення експериментального середовища. Після проведення експертної оцінки серед 50 спеціалістів у сфері аналізу даних з України, Польщі, Австрії, Португалії та США ключовими критеріями ефективності були обрані такі показники:

- точність, визначена за допомогою показників Precision і Recall з коефіцієнтом важливості 16;
- економія часу цільового алгоритму з коефіцієнтом важливості 2;
- збереження мінімально допустимої кількості цільових даних для досягнення «точності» більше 80% з коефіцієнтом важливості 2.

Що стосується наборів даних, було вирішено обрати новини присвячені повномасштабному вторгненню Росії в Україну та поширення неправдивих відомостей під час виборчого процесу 2020 року в США. Описані набори даних були сформовані авторами роботи за допомогою соціальної мережі X.

Результати проведеного експерименту показали, що точність запропонованого підходу приблизно на 17% вища за точність базового рішення на основі нейронних мереж. Не зважаючи на те, що швидкодія простого RCNN більша за двошаровий підхід, рівень збереження мінімально допустимої кількості даних для досягнення «точності» більше 80%, дозволяє досягнути ефективності, яка суттєво більша за модель для порівняння.

Дослідження мало на меті створити двошарову модель класифікації даних, яка, окрім визначення факту фабрикації, могла б також оцінити ступінь і мету фальсифікації даних. Результати показали, що підхід із використанням двошарової моделі досягав у середньому 15% покращення ефективності порівняно з прямим використанням методу RCNN. Цей висновок підкреслює потенціал гібридного підходу для використання в системах з високим навантаженням, де швидка ідентифікація та відповідь на сфабриковану інформацію є вирішальними.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. E. Aïmeur, S. Amri and G. Brassard, «Fake news, disinformation and misinformation in social media: a review», *Social Network Analysis and Mining*, vol. 13, article 30, 2023. Accessed: Nov 8, 2024. [Online]. Available: <https://doi.org/10.1007/s13278-023-0102>

2. D. D. Giandomenico, J. Sit, A. Ishizaka and D. Nunan, «Fake news, social media and marketing: A systematic review», *Journal of Business Research*, vol. 124, pp. 329–341, 2021. Accessed: Nov 8, 2024. [Online]. Available: <https://doi.org/10.1016/j.jbusres.2020.11.037>.

3. S. Yakovlev, A. Khovrat, V. Kobziev and D. Uzlov, «Decision Support Algorithm in the Development of Information Sensitive Socially Oriented Systems», in *Profit AI*, Cambridge, USA, Sep. 25–27, 2027. CEUR Workshop Proceedings, 2024, pp. 315–326. Accessed: Nov 8, 2024. [Online]. Available: <https://ceur-ws.org/Vol-3777/paper20.pdf>.

4. S. Yakovlev, A. Khovrat and V. Kobziev, «Using Parallelized Neural Networks to Detect Falsified Audio Information in Socially Oriented Systems», in *Information Technology and Implementation*, Kyiv, Ukraine, Nov. 20–21, 2023. CEUR Workshop Proceedings, 2023, pp. 220–238. Accessed: Nov 8, 2024. [Online]. Available: https://ceur-ws.org/Vol-3624/Paper_19.pdf.

5. A. Choudhary and A. Arora, «Linguistic feature based learning model for fake news detection and classification», *Expert Systems with Applications*, vol. 169, article 114171, 2021. Accessed: Nov 8, 2024. [Online]. Available: <https://doi.org/10.1016/j.eswa.2020.114171>.