УДК 004.021 DOI: 10.18372/2073-4751.73.17643

Міщенко Л.Д.,
Клименко І.А., д.т.н.,
orcid.org/0000-0001-5345-8806

СПОСІБ ПРИСКОРЕНОГО РОЗПІЗНАВАННЯ ФЕЙКОВИХ НОВИН НА ОСНОВІ ОБРОБКИ ПРИРОДНОЇ МОВИ ТА ВИДАЛЕННЯ ГОЛОСНИХ ЛІТЕР У СЛОВАХ

Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"

mishchenko.liudmyla@lll.kpi.ua

Вступ

Використання різноманітних новинних веб-ресурсів в Інтернеті та соціальних мережах для поширення інформації стає все більш популярним. Здобуваючи певну читацьку аудиторію та завойовуючи її довіру, такі джерела починають поширювати фейкові новини чи маніпуляції. Тому ідея захисту населення від дезінформації та поширення маніпулятивного впливу під час війни є надзвичайно гострою та необхідною сьогодні.

Використання сучасних технологій є необхідним чинником у боротьбі з поширенням фейкових даних. Крім того, головним завданням є швидкий автоматичний аналіз інформації, а також поширення спростувань і правдивих фактів. Тому розробка нових алгоритмів пошуку та аналізу щоденного потоку новин є надзвичайно актуальним завданням.

Важливою частиною впровадження будь-якого програмного забезпечення є його ефективність і швидкість. Поширення фейкової інформації можна спостерігати щодня, але автоматичної перевірки інформації практично немає. Всі перевірки новин здійснюються виключно журналістами або автоматизованими системами, які мають вкрай обмежений і недостатньо швидкий функціонал. Тому пропонується метод підвищення ефективності валідації новин шляхом використання технології Natural Language Processing (NLP) разом з алгоритмом Левенштейна.

На даний момент вивчено та описано значну кількість проблем у сферах IT, медицини, економіки, кримінології та інших, які можуть бути вирішені за допомогою технології розпізнавання мовлення. Проте основна частина публікацій спрямована на застосування аналізу природного мовлення до різних професійних сфер. Є роботи, в яких описано використання NLP для аналізу настроїв, відстеження та контролю громадської думки, пошуку текстів на певну тему та виділення ключових слів, аналізу пошукових систем, використання чат-ботів.

Сьогодні гостро стоїть питання ефективного аналізу інформації, в тому числі й новинної, автоматичними рішеннями. Однак використання технології NLP для ефективної обробки та вилучення фейкових новин залишається невивченим.

Мета

Метою дослідження є розробка ефективної системи для виявлення, синтезу та аналізу новин за допомогою технології обробки природної мови. При цьому, у запропонованому рішенні розглядається спосіб прискорено аналізу тексту на основі базового алгоритму [1] із застосуванням скорочення слів за рахунок видалення голосних літер у них для значного зменшення ваги векторного представлення слів та прискорення їх аналізу.

Основна частина

У вже існуючих рішеннях, автори зазвичай розглядають рішення аналізу тексту, проте також є роботи у сфері аналізу зображення.

Одне з великих досліджень провела група дослідників у сфері соціальних

мереж як основної платформи для поширення місінформації. Вони досліджують нову проблему використання соціального контексту для виявлення фейкових новин. Запропонували структуру вбудовування трьох зв'язків TriFN, яка одночасно моделює відносини між видавцем і новинами та взаємодію між новинами та користувачем для класифікації фейкових новин. Провели експерименти на двох реальних наборах даних, які демонструють, що запропонований підхід значно перевершує інші базові методи виявлення фейкових новин [2].

Іще важливе дослідження у цьому напрямку поєднує три загальноприйняті характеристики фейкових новин: текст статті, відповідь користувача, яку вона отримує, і джерело користувачів, які її рекламують. Бо існуючі роботи здебільшого зосереджені на пристосуванні рішень до однієї конкретної характеристики, що обмежує їх успіх і загальність. Проте автори у дослідженні [3] запропонували модель, яка поєднує всі три характеристики для більш точного та автоматизованого прогнозу. Зокрема, враховують поведінку обох сторін, користувачів і статей, а також групову поведінку користувачів, які свідомо чи ні поширюють фейкові новини.

Невирішена проблематика в більшості існуючих алгоритмів виявлення фейкової інформації полягає в тому, що вони зосереджені на пошуку підказок у наповненні новин, які, як правило, неефективні, оскільки фейкові новини часто навмисно пишуться, щоб ввести користувачів в оману шляхом імітації правдивих новин. Значна частина новин активно поширюється користувачами соцмереж і через надмірну популярність системи не сприймають їх як неправду чи маніпуляцію.

Іще однією невирішеною проблемою є значне збільшення об'єму нового контенту у соціальних мережах та новинних платформах. Навіть автоматизовані рішення неспроможні впоратися з якісним аналізом. Оскільки навчання нейронних мереж на базі нового контенту відбувається повільніше, ніж створення фейків чи дезінформації. Тобто важливо ідентифікувати фейки на ранній стадії поширення.

Також дослідження цієї теми зазвичай відбувається з аналізом соціальних мереж та трендів, проте алгоритми для аналізу новинних сайтів із величезною аудиторією залишаються осторонь напрацювань.

Однією з важливих етапів опрацювання вхідного тексту є швидкість його аналізу. Для швидкої обробки тексту необхідно зробити його з ефективною обчислювальною архітектурою даних. Саме це стоїть як необхідне завдання для дослідження в запропонованій публікації та має бути виконаним наступним кроком.

Для побудови ефективної архітектури даних використовують Skip-gram модель. Ця модель має на меті навчити просту нейронну мережу з одним прихованим шаром для виконання певного завдання. При цьому, важливим етапом розробки мережі є дослідження значення ваги прихованого шару, яке в дійсності відповідає "векторам слів", з якими відбувається навчання.

Вхідними даними для такої мережі може бути велика кількість неструктурованих текстових даних, з яких формується словник із унікальних слів. При цьому, кожне слово представляється як одноразовий (one-hot) вектор.

Результатом роботи мережі буде єдиний вектор, який містить для кожного слова у раніше підготовленому словнику ймовірність того, що випадково вибране сусіднє слово є цим словом зі словника.

Під час навчання такої мережі на парах слів вхідним сигналом є одноразовий вектор, що представляє вхідне слово, а навчальним виходом також є одноразовий вектор, що представляє вихідне слово. Але для того, щоб оцінити уже навчену мережу за вхідним словом, потрібно звертатися за результатом до вихідного вектору. Він буде представлений розподілом імовірностей, тобто набором значень з плаваючою комою, а не одноразовим вектором.

Отже, узагальненою метою цього способу є лише обрахунок значення вагів

прихованого шару матриці. При цьому, важливо пам'ятати, що нейронна мережа нічого не знає про зсув у шарах навчання та про те, що вихідне значення є відносний, а не абсолютним результатом. Таким чином, якщо початкова вибірка матиме занадто малу кількість слів, тоді кінцевий результат не буде наближеним до 100% [4].

У вже готових рішеннях на основі Skip-gram моделі є реалізовані рішення для таких завдань, як складання аналогій, наприклад, «Німеччина» : «Берлін» :: «Франція» : X. Вони розв'язуються шляхом знаходження такого вектора x, при якому vec(x) є найближчим до vec(«Берлін») - vec(«Німеччина» ") + vec("Франція") відповідно до косинусної відстані (викидаємо введені слова з пошуку). Цей конкретний приклад вважається правильною відповіддю, якщо X є «Париж». При цьому, завдання має дві великі категорії: синтаксичні аналогії (наприклад, «швидко» : «швидко» :: «повільно» : «повільно») і семантичні аналогії, такі як зв'язок між країною та столицею.

Одним зі способів рішення є додавання векторів слів. Це продемонстровано на базі моделі Skip-gram, де представлення слів і фраз демонструють лінійну структуру. Це дає змогу виконувати точні задачі на аналогію за допомогою простої векторної арифметики. Але також є вагомим спосіб поєднання слів шляхом поелементного додавання їхніх векторних представлень. Адитивну властивість векторів можна пояснити шляхом перевірки мети навчання. Вектори слів перебувають у лінійній залежності від вхідних даних для нелінійності softmax. Оскільки вектори слів навчені передбачати навколишні слова в реченні, вектори можна розглядати як представлення розподілу контексту, у якому з'являється слово. Ці значення логарифмічно пов'язані з ймовірностями, обчисленими вихідним рівнем, тому сума двох векторів слів пов'язана з добутком двох контекстних розподілів. Продукт працює тут як функція I: слова, яким обидва вектори слів присвоюють високу ймовірність, матимуть високу ймовірність, а інші слова матимуть низьку ймовірність [5]. Таким чином, якщо «річка Дніпро» часто зустрічається в одному реченні разом зі словами «українська» і «річка», сума цих двох векторів слів призведе до такого вектора ознак, який близький до вектора «річка Дніпро» .

Іще одним варіантом рішення є спосіб обчислення векторів фраз, замість векторів окремих слів. Про це у своєму дослідженні описують автори, які намагаються знаходити слова, які часто зустрічаються разом і рідко в інших контекстах. Наприклад, "Wall Street Journal" та «New York Times» презентовано унікальними токенами в навчальних даних. З точки зору тематики дослідження, такі фрази мають значно вагоміше значення, а ніж окремі слова як одиниці тексту. Адже вони частіше зустрічаються як сталий вираз у сфері новин.

Для отримання прискорення в обробці тексту, зазвичай необхідно зменшити кількість вхідних даних для навчання моделі. При цьому, можна розглянути два типи навчання моделей: онлайн-навчальну та попередньо навчену моделі. Для моделі із онлайн навчання використовується вхідний набір даних як навчальні дані, на основі яких генерується словниковий запас та вектори слів відповідно. Попередньо навчена модель зазвичай базується на значно більшому обсязі текстових даних. Тобто для навчання може використовуватися величезна кількість вхідних даних, наприклад, збірка усіх новин з різних ресурсів за декілька років. Але результатом попередньо підготовленої моделі буде набір пар слів або вбудовування (collection of word / embedding pairs). Так попередньо навчена модель узагальнює словниковий запас із вхідного набору даних і генерує вектор вбудовування для кожного слова з попередньо навченої моделі. Без онлайн-навчання використання попередньо навченої моделі може заощадити час навчання. Він має кращу продуктивність, особливо коли розмір вхідного набору даних відносно малий [6].

Таким чином, для запропонованого способу обрано попередньо навчену модель. Адже є завідома підготовлена база перевірених новин, яка не досягає обсягів даних, наприклад, Вікіпедії чи Твітера. Тому важливо мати більш точно навчену модель.

Запропонований спосіб полягає в тому, щоб зробити обробку вхідного набору слів безпосередньо до початку попереднього навчання. Це реалізується для того, аби зменшити розмір і кількість слів, що в подальшому дозволить моделі навчитися скоріше.

Реалізація відбувається шляхом видалення голосних літер у словах після етапу лемітизації. Завдяки цьому отримано слова меншого розміру, а також під час етапу перетворення слів на токени, буде зменшено кількість токенів. Це дає можливість уникнути великої кількості нулів у векторному представлені словника слів та утворенню так званих розряджених векторів (sparse vectors), відповідно досягти мінімальної обчислювальної складності тексту.

У основі архітектури дослідження взято Skip-gram модель та обмежену коротку довжину вбудованого вектора слів.

Загалом довжина вбудовування слів становить кілька сотень. Наприклад, 300, 400, 700. Хоча нормальним вважається теж, якщо довжина вимірюється тисячею й більше. Зазвичай більше тисячі зустрічається у випадку із вбудованими словосполученнями.

Невеликий розмір вбудовування означає малий векторний простір, таким чином швидший аналіз слів. Проте при значному скороченні слова, наприклад, як у запропонованому способі, векторний простір може зменшитися до таких розмірів, що є висока ймовірність спричинення колізій під час вбудовування слів.

Для досягнення найшвидшої обробки тексту, запропоновано фіксувати довжину вбудованих слів для попередньо навчених моделей. При цьому, мінімальне значення становить 100, що дозволяє уникнути значної кількості колізій. Максимальне значення – 300. З таким показником навчання моделі не буде мати затримок у швидкості.

Висновки

У запропонованій стаття розглянуто спосіб розпізнавання фейкових новин у мережі Інтернет, базуючись на технології Natural Language Processing і алгоритму Левенштейна. При цьому, використовується архітектура Skip-gram model та виконано додатковий крок підготовки тексту до аналізу – вилучення голосних літер зі слів. Цей процес відбувається після проходження етапу лемітизації NLP. Він необхідний для зменшення розміру вбудованих слів-токенів, а в подальшому, меншого значення векторного представлення слова. Завдяки такій додатковій обробці слів, вдалося зменшити затримки при обчислені та порівнянні довгих розряджених векторів.

Оскільки, у запропонованому способі використовуються короткі векторні значення, як 100 чи 300, то є ймовірність виникнення колізій. Коли векторне представлення одного слова, дорівнює значенню вектора іншого. Проте ці колізі трапляються доволі рідко, тому ця частина залишається відкритою для наступних досліджень.

Таким чином, можна вважати, що мета – прискорити систему виявлення, синтезу та аналізу фейкових новин за рахунок використання технології Natural Language Processing із попередньою обробкою слів, за рахунок їх скорочення, – досягнута.

Запропонований спосіб реалізації поставленого завдання виявився ефективним для розпізнавання фейкових новин з досить високою точністю та виявився значно швидшим за існуючі рішення.

Література

1. *Mishchenko L., Klymenko I.* Method for detecting fake news based on natural language processing. The VI International Scientific and Practical Conference “Modern ways of solving the problems in science in the world”, February 13-15, Warsaw, Poland. – P. 375-378. URL: https://eu-conf.com

/ua/events/modern-ways-of-solving-the-problems-of-science-in-the-world/.

2. *Zhou X., Zafarani R.* A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys (CSUR), 2020. – Vol. 53 – No. 5. – P. 1-40.

3. *Ruchansky N., Seo S., Liu Y.* CSI: A hybrid deep model for fake news detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. – 2017. – P. 797-806.

4. *McCormick Ch.* Word2Vec Tutorial – The Skip-Gram Model. – P. 1-5. URL: https://www.fer.unizg.hr/_download/repository/TAR-2020-reading-05.pdf.

5. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* Distributed Representations of Words and Phrases and their Compositionality. Advances in neural information processing systems, 2013, – P. 26. URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

6. Convert Word to Vector component. Microsoft documentation, 2021. URL: https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/convert-word-to-vector.

Міщенко Л.Д., Клименко І.А.

СПОСІБ ПРИСКОРЕНОГО РОЗПІЗНАВАННЯ ФЕЙКОВИХ НОВИН НА ОСНОВІ ОБРОБКИ ПРИРОДНОЇ МОВИ ТА ВИДАЛЕННЯ ГОЛОСНИХ ЛІТЕР У СЛОВАХ

Новинні веб-ресурси набувають більшої популярності в наші дні. Такі джерела інформації можуть використовувати довіру аудиторії для маніпулювання фактами та поширення фейків. Таким чином, захист від таких ресурсів є величезною проблемою сьогодення.

Найважливішою частиною будь-якого програмного забезпечення є його швидкість роботи. Фейки з'являються щодня, але сьогодні немає систем автоматичної перевірки фактів. Усі перевірки виконуються журналістами або напів автоматизованими системами, які або специфічні для невеликих завдань, або занадто повільні. Тому ця стаття пропонує спосіб перевірки фактів за допомогою NLP та алгоритму Левенштейна. При цьому, у способі запропоновано прискорений аналіз тексту, роблячи обчислення з мінімальним значенням векторного представлення слів. Це вдалося досягти на рівні лемітизації NLP за рахунок видалення голосних літер зі слів.

У наші часи, тема вивчена досить глибоко. Але більшість досліджень зосереджені на використанні технології NLP для природного аналізу мовлення в конкретних галузях, таких як пошук тексту, боти, розмітка тексту тощо. До того ж, не розглядалося зменшення векторного представлення слів для прискореного аналізу тексту та структурування його токенів.

Основне завдання дослідження полягає в розробці ефективної системи виявлення підробок за допомогою технології Natural Language Processing, яка показує результат доволі швидко за рахунок зменшення довжини слів, а не базуючись на попередньому навчанні системи.

У роботі доведено здатність технології NLP вирішувати завдання перевірки фактів. Проте ще є кілька напрямків для подальшої роботи. Наприклад, використання навчальної нейронної мережі для виявлення найбільш розповсюджених підробок або дослідження виникнення колізій у векторному представленні коротких слів без голосних літер.

***Ключові слова**: технологія Natural Language Processing, фейк, маніпуляція, аналіз тексту, відстань Левенштейна, векторне представлення слова.*

Mishchenko L.D., Klymenko I.A.

A METHOD OF ACCELERATED FAKE NEWS RECOGNITION BASED ON NATURAL LANGUAGE PROCESSING AND REMOVAL OF VOWELS IN WORDS

News web resources are gaining more popularity these days. Such sources of information can use the trust of the audience to manipulate facts and spread fakes. Thus, protection against such resources is a huge challenge today.

The most important part of any software is its speed. Fakes appear every day, but today there are no automatic fact-checking systems. All checks are done by journalists or by semi-automated systems that are either specific to small tasks or too slow. Therefore, this article proposes a fact-checking method using NLP and Levenshtein algorithm. At the same time, the method offers accelerated text analysis, making calculations with the minimum value of the vector representation of words. This was achieved at the level of NLP lemmatization by removing vowels from words.

In our time, the topic is studied quite deeply. But most research focuses on using NLP technology for natural speech analysis in specific fields such as text search, bots, text markup, etc. In addition, the reduction of the vector representation of words for accelerated text analysis and structuring of its tokens was not considered.

The main task of the research is to develop an effective forgery detection system using Natural Language Processing technology, which shows the result quite quickly by reducing the length of words, and not based on the previous training of the system.

The paper proves the ability of NLP technology to solve the task of fact checking. However, there are still several directions for further work. For example, using a learning neural network to detect the most common forgeries or investigating the occurrence of collisions in the vector representation of words without vowels.

***Key words**: Natural Language Processing technology, fake, manipulation, text analysis, Levenshtein distance, vector word representation.*