https://doi.org/10.31649/1997-9266-2024-177-6-121-128

УДК 004.9+007.76

Б. С. Білецький[1]
В. Б. Мокін[1]

ВИЗНАЧЕННЯ ТЕМПОРАЛЬНОЇ СПРЯМОВАНОСТІ В ТЕКСТАХ: НЕЙРОМЕРЕЖЕВИЙ ПІДХІД ДЛЯ ХРОНОЛОГІЧНОГО ВПОРЯДКУВАННЯ НА ОСНОВІ АНАЛІЗУ ПАР СЛІВ

[1]Вінницький національний технічний університет

Запропоновано нейромережевий підхід до визначення темпоральної спрямованості у текстах, що дозволяє відтворювати хронологію подій, навіть за відсутності явних часових маркерів. Цей підхід визначає ймовірнісний порядок появи слів у текстах з урахуванням їхніх статистичних та лінгвістичних зв'язків. На відміну від традиційних підходів, які покладаються на явні часові вирази або дати публікацій, запропонований підхід дає змогу оцінювати порядок подій на основі виявлених взаємозв'язків між парами слів в документах, що описують події.

Для аналізу темпоральної спрямованості використовуються нейронні мережі, що дозволяють моделювати відносини між словами шляхом попарного порівняння їхньої появи в текстах. Запропоновано формули для обчислення показників темпоральної спрямованості, які базуються на частоті появи слів у датованих текстах. Отримані показники нормалізовані, що забезпечує кращу інтерпретацію результатів.

На основі цих показників сформовано набір ознак для тренування моделей машинного навчання за різними критеріями. Для перевірки ефективності створено україномовний корпус із 127 000 новин соціальних мереж та застосовано кілька моделей: Gradient Boosting Classifier, Random Forest Classifier, Decision Tree та Logistic Regression. Як приклад, вибрано 48 ознак, які характеризують ці новини. У ході експериментів виявлено, що модель Gradient Boosting Classifier показала найкращий результат з точністю 89,76 % на валідаційному датасеті, що перевищило точність інших моделей, таких як Random Forest (74,81 %) та Decision Tree (68,97 %).

Запропонований підхід підтвердив ефективність у моделюванні хронологічних зв'язків між подіями, що є важливим для задач автоматизації текстів. Підхід можна використовувати для аналізу новин, хронологічного впорядкування історичних подій і роботи з текстовими даними у великих масивах.

Ключові слова: інтелектуальні технології, машинне навчання, штучний інтелект, нейронні мережі, оброблення природної мови, темпоральна спрямованість, інформаційна технологія.

Вступ

Темпоральне впорядкування подій у текстах є важливим завданням обробки природної мови (англ.: «Natural Language Processing», скорочено: «NLP»). Це завдання широко використовується в таких галузях, як автоматизація новин, аналіз історичних текстів і хронологічне впорядкування подій у документальних джерелах. Часові зв'язки між подіями дають змогу зрозуміти їхню послідовність, причинно-наслідкові зв'язки й вплив на інші події, що робить це завдання надзвичайно важливим для створення точних і функціональних систем аналізу текстових даних. Однак багато текстів, зокрема новини, часто не містять чітких часових міток, що ускладнює встановлення точного хронологічного порядку подій.

Різні підходи до темпорального впорядкування подій нині активно розвиваються [1]. У деяких роботах пропонуються моделі, що використовують синтаксично-керовані мережі граф-трансформаторів для вилучення часових зв'язків між подіями [2], [3]. Ballesteros et al., Ning [4], [5] та Goyal і Durrett [6] пропонують застосовувати нейронні мережі та методи векторного представлення слів для визначення темпоральних зв'язків між подіями на основі часових виразів та їхніх взаємозв'язків. Інші дослідження, такі як Liu et al. [7] та Xia et al. [8], зосереджуються на моделюванні

темпоральних взаємодій у графових структурах, прогнозуючи час зв'язку між вузлами в динамічних мережах. Q. Ning та ін. запропонували вдосконалену нейромережеву модель для визначення часових зв'язків між подіями, приділяючи увагу структурному представленню часових залежностей [9]. A. Naik зі співавторами запропонували корпус TDDiscourse, що дає змогу аналізувати часові зв'язки між подіями на рівні дискурсу з урахуванням віддаленості подій у тексті [10]. Проте більшість з цих досліджень або працюють з явними часовими маркерами (часовими виразами, датами), або зосереджуються на подіях у текстах, не враховуючи контексту та темпоральних зв'язків між словами.

Однією з основних проблем є встановлення точного порядку подій у текстах, де часові мітки або відсутні, або є надто загальними. Наявні підходи часто не враховують лінгвістичних моделей, які можуть вказувати на темпоральну спрямованість подій на рівні слів. Наприклад, слова, що мають контекстні зв'язки, такі як «злочин» і «покарання» або «розроблення» і «впровадження», можуть допомогти передбачити хронологію подій навіть без явного зазначення часу.

Метою цього дослідження є розробка нейромережевого підходу для визначення темпоральної спрямованості між словами в текстах новин з метою відтворення хронології подій. На відміну від попередніх досліджень, наш підхід зосереджується на попарному порівнянні слів у новинних текстах задля аналізу темпоральних взаємозв'язків між подіями. Пропонується підхід до аналізу текстів, який використовує нейронні мережі для моделювання ймовірності появи одного слова перед іншим. Це дасть змогу точніше визначати хронологічну послідовність подій навіть у тих текстах, де відсутні часові маркери.

Для цього використовується великий корпус новин із соцмереж, який дає змогу працювати з автентичними текстами, де події часто подаються без чітких часових міток. Очищення текстів і лематизація (процес приведення слів до їхньої базової форми [6]) дають змогу створити словник унікальних слів, що використовується для попарного порівняння та оцінювання темпоральної спрямованості між словами. Основою запропонованого підходу є формула для обчислення показника темпоральної спрямованості, що базується на аналізі датованих новин.

Формалізація задачі та ідея її розв'язання

Розглянемо датасет з n текстів новин (документів) у межах одного контексту $\mathrm{D} = \{d_1, d_2, \ldots d_n\}$. Серед них є множина документів $D_{\text{виз}}$, для яких відомий час публікації $t(d_n)$, також існує множина документів $D_{\text{невиз}}$, для яких час публікації невідомий або невизначений.

Текст новини може містити опис кількох подій, що відбулися в різний час, або представлення подій, які мають внутрішню хронологічну структуру, незалежну від дати публікації. Це зумовлює необхідність розв'язання таких задач:

– автоматично визначати відносний порядок появи подій у текстах новин без явних часових позначок. Використання лише дати публікації є недостатнім для точного встановлення порядку подій, особливо якщо в одному документі описується кілька подій або події висвітлюються в різних контекстах;

– оцінювати темпоральну спрямованість між словами, що відображають послідовність подій, на основі лінгвістичних та статистичних закономірностей. Це дає змогу глибше аналізувати внутрішню структуру тексту та встановлювати причинно-наслідкові зв'язки між подіями, навіть за відсутності явних часових маркерів;

– забезпечувати високу точність у відтворенні хронології документів на основі аналізу пар слів у текстах.

Конкретно, головне завдання полягає у розробленні нейромережевого підходу, який:

– моделює ймовірність того, що одне слово з'являється раніше за інше, шляхом попарного порівняння слів у новинних текстах. Це дає змогу враховувати контекстні зв'язки між словами та встановлювати точнішу хронологічну послідовність подій всередині тексту;

– агрегує оцінки темпоральної спрямованості на рівні слів для визначення загального хронологічного порядку документів;

– демонструє високу точність у хронологічному впорядкуванні документів навіть за відсутності явних часових маркерів. Це особливо важливо для автоматичного аналізу великих масивів текстових даних, де ручне визначення порядку подій є непрактичним.

Таким чином, *постановка задачі* полягає в розробленні нейромережевого підходу, здатного ви-

значати темпоральну спрямованість між словами в текстах новин без явних часових маркерів, та демонстрації його ефективності на великому корпусі даних.

Це дасть змогу не лише автоматизувати процес хронологічного впорядкування новинних подій, а й підвищити точність аналізу великих масивів текстових даних, що є важливим для таких застосувань, як аналіз історичних подій, моніторинг медіа, встановлення причинно-наслідкових зв'язків між подіями, а також інших сфер, де необхідно розуміти послідовність.

Для відтворення хронології подій у першій множині $D_{\text{виз}}$ достатньо їх відсортувати. Щодо другої множини документів $D_{\text{невиз}}$, то можна передбачити хронологію подій, оцінюючи їхню темпоральну спрямованість.

Темпоральна спрямованість (Temporal Directionality), запропонована в цьому дослідженні, суттєво відрізняється від відомих понять у сфері темпоральної інформації та обробки природної мови.

По-перше, на відміну від темпоральних відносин (наприклад, «до», «після», «під час»), що визначають логічні або хронологічні зв'язки між подіями чи часовими виразами, темпоральна спрямованість ґрунтується на ймовірнісному підході для оцінювання порядку появи слів з огляду на статистичні закономірності в текстових даних.

По-друге, традиційне темпоральне впорядкування передбачає чітку хронологічну послідовність подій, засновану на часових маркерах або контексті. Натомість темпоральна спрямованість дає змогу визначати відносний порядок подій навіть за відсутності таких маркерів, аналізуючи взаємну частоту появи слів у датованих текстах.

По-третє, на відміну від концепції «темпоральної прив'язки» (temporal anchoring), що зосереджується на присвоєнні подіям абсолютних часових позначень, темпоральна спрямованість орієнтована на відносні зв'язки між словами, даючи змогу моделювати гнучкіші й адаптивніші відносини без необхідності точного визначення часових рамок.

До того ж, «темпоральне міркування» (temporal reasoning) включає процеси логічного виведення темпоральних зв'язків на основі попередньо заданих даних. Водночас темпоральна спрямованість фокусується на безпосередньому визначенні ймовірності порядку слів без опори на попередні логічні висновки щодо відносин.

Отже, нейромережевий підхід визначення темпоральної спрямованості міг би доповнити наявні підходи темпорального аналізу, дозволяючи оцінювати відносний порядок подій на основі статистичних даних, навіть за відсутності явних часових маркерів. Це розширило б можливості автоматичного аналізу текстів та дозволило підвищити точність хронологічного впорядкування подій у різноманітних контекстах.

Визначення темпоральної спрямованості між двома документами пропонується здійснювати у два етапи:

Eman 1. Визначення темпоральної спрямованості між усіма парами слів цих документів, які несуть певний унікальний зміст (за винятком стоп-слів, спеціальних символів, слів, які зустрічаються в обох документах одночасно, тощо), у вигляді числової оцінки в діапазоні [0, 1] або у нормалізованому вигляді [–1, 1].

Eman 2. Обчислення різних статистичних показників щодо слів і символів у цих документах та аналіз темпоральної спрямованості пар слів у цих документах (наприклад, порівняння з різними порогами допустимого діапазону) та подальша класифікація за одним з 3-х варіантів: перший документ раніше другого, другий — раніше першого чи «Не можна це визначити» (там може й не бути часової прив'язки, наприклад «Всесвіт — нескінченний»).

Формалізуємо задачу. Темпоральна спрямованість TD_d між парою документів d_1 та d_2 — це ймовірність того, що публікація документа d_1 відбулася раніше, ніж d_2.

Визначити темпоральну спрямованість на рівні документів можна, оцінюючи сукупність значень показника темпоральної спрямованості між парами слів. Для цього необхідно попарно порівняти всі слова $w_i \in d_1$ з усіма словами $w_j \in d_2$.

Показник темпоральної спрямованості (ПТС) для пари слів можна обчислити, маючи датасет датованих новин, що містить тексти новин та дати їхніх публікацій за формулою $\mathrm{TD}\left(w_i, w_j\right)$.

Формула для обчислення показників темпоральної спрямованості між словами $w_i \in d_1$ та $w_j \in d_2$ пропонується у такому вигляді:

$$TD\left(w_i, w_j\right) = \frac{N\left(w_i \to w_j\right)}{N\left(w_i\right) \bullet N\left(w_j\right)},\ i = \overline{1,n},\ j = \overline{1,n}, \tag{1}$$

де $N\left(w_i \to w_j\right)$ — кількість випадків, коли документ, що містить слово w_i, опубліковано раніше за документ, що містить слово w_j; $N\left(w_i\right)$ — загальна кількість документів, що містять слово w_i; $N\left(w_j\right)$ — загальна кількість документів, що містять слово w_j.

Ця формула обчислює ймовірність того, що слово w_i з'являється в публікаціях раніше, ніж слово w_m, на основі дат публікацій у датасеті новин:

– якщо $TD\left(w_i, w_j\right) = 0$, тоді слово w_i не з'являлося ніколи перед w_j, а отже, з'являлось завжди після w_j;

– якщо $TD\left(w_i, w_j\right) = 0{,}5$, тоді слово w_i з'являється як раніше, так і пізніше w_j;

– якщо $TD\left(w_i, w_j\right) = 1$, тоді слово w_i завжди зустрічається після w_j.

Для наочності можна запропонувати нормалізований у межах (–1, 1) варіант цього показника $NTD\left(w_i, w_j\right)$

$$NTD\left(w_i, w_j\right) = 2 \cdot TD\left(w_i, w_j\right) - 1. \tag{2}$$

Ця формула обчислює нормалізоване значення темпоральної спрямованості між словами w_i та w_j, де

– якщо $NTD\left(w_i, w_j\right) = -1$, це означає, що слово w_i завжди з'являється після w_j;

– якщо $NTD\left(w_i, w_j\right) = 0$, це означає, що обидва слова w_i та w_j зустрічаються в текстах приблизно одночасно, без чіткої хронологічної переваги одного над іншим;

– якщо $NTD\left(w_i, w_j\right) = 1$, це означає, що слово w_i завжди з'являється перед w_j.

У результаті обчислення показників темпоральної спрямованості на рівні слів $TD\left(w_i, w_j\right)$ для кожної пари слів $w_i \in W_1$ та $w_j \in W_2$ між документами d_1 та d_2, утворюється така сукупність:

$$NTD_{words}\left(W_1, W_2\right) = \begin{pmatrix} ntd_{11} & \cdots & ntd_{1m} \\ \vdots & \ddots & \vdots \\ ntd_{n1} & \cdots & ntd_{nm} \end{pmatrix}, \tag{3}$$

де W_1, W_2 — це сукупності слів у документах d_1 та d_2, $ntd_{nm} = NTD\left(w_i, w_j\right)$ — це значення показника темпоральної спрямованості між парою слів w_i та w_j.

Для кожної пари документів матриця буде розміру $A \cdot B$ де A — кількість слів у документі d_1, а B — кількість слів у документі d_2. Оскільки в цьому дослідженні не враховується положення слів у тексті, а кількість слів у документах може бути різною, показник $NTD_{words}\left(W_1, W_2\right)$ можна представити у вигляді вектора довжиною $A \cdot B$.

Важливо зазначити, що кількість операцій, необхідних для обчислення темпоральної спрямованості між двома документами $TD()$ та $NTD()$, буде пропорційна: $O\left(A \cdot B \cdot k^2\right)$, де k — середня кількість документів, в яких зустрічається кожне слово. Таким чином, складність алгоритму зростає лінійно зі збільшенням кількості слів у документах та квадратично зі збільшенням кількості документів, в яких зустрічаються ці слова. На датасеті з 150 000 документів по 50 слів у кожному кількість операцій може сягати $5{,}6 \cdot 10^9$, що ускладнює створення великого набору даних для навчання на комп'ютерах з обмеженими ресурсами та недостатньою потужністю. А тому для оброблен-

ня цих даних пропонується застосування глибоких нейронних мереж з паралелізацією обчислень.

На другому етапі для класифікації темпоральної спрямованості документів J варто здійснювати передбачення за різними ознаками:

$$F(W_1), F(W_2), G(NTD_{words}), \ldots, \quad (4)$$

де $F()$ — ознаки, основані на статистичних показниках, обчислених для вибірок кількості слів та символів у документах, $G()$ — ознаки, які дозволяють по темпоральній спрямованості (*NTD*) між словами оцінити спрямованість між документами та взаємозв'язки між ними.

До прикладу, для (4) можна запропонувати набір з 48 ознак (їх може бути й набагато більше), поданих у табл. 1. В процесі тривалих експериментальних досліджень встановлено, що саме цей набір ознак забезпечує найкращу продуктивність моделі, у порівнянні з іншими варіантами.

Таблиця 1

Перелік ознак для аналізу слів у документах

№	Опис ознаки	№	Опис ознаки
1	Кількість слів у тексті 1 (W_1)	25	Сума всіх знаків
2	Кількість слів у тексті 2(W_2)	26	Кількість оцінок *NTD*, які менше –1
3	Кількість символів у тексту 1 (W_1)	27	Кількість оцінок *NTD*, яка дорівнює –1
4	Кількість символів у тексті 2(W_2)	28	Кількість оцінок, яка дорівнює –0,9
5	Добуток кількості слів у текстах 1(W_1) і 2(W_2)	29	Кількість оцінок *NTD*, яка дорівнює –0,8
6	Сума кількості слів у текстах 1(W_1) і 2(W_2)	30	Кількість оцінок *NTD*, яка дорівнює –0,7
7	Відношення кількості слів у тексті 1(W_1) до тексту 2(W_2)	31	Кількість оцінок *NTD*, яка дорівнює –0,6
8	Відношення кількості слів у тексті 2(W_2) до тексту 1(W_1)	32	Кількість оцінок *NTD*, яка дорівнює –0,5
9	Добуток кількостей символів текстів 1(W_1) і 2(W_2)	33	Кількість оцінок *NTD*, яка дорівнює –0,4
10	Сума кількостей символів текстів 1(W_1) і 2(W_2)	34	Кількість оцінок *NTD*, яка дорівнює –0,3
11	Відношення кількостей символів тексту 1(W_1) до тексту 2(W_2)	35	Кількість оцінок *NTD*, яка дорівнює –0,2
12	Відношення кількостей символів тексту 2(W_2) до тексту 1(W_1)	36	Кількість оцінок *NTD*, яка дорівнює –0,1
13	Середнє значення оцінок *NTD*	37	Кількість оцінок *NTD*, яка дорівнює 0
14	Знак середнього значення оцінок *NTD*	38	Кількість оцінок *NTD*, яка дорівнює 0,1
15	Стандартне відхилення оцінок *NTD*	39	Кількість оцінок *NTD*, яка дорівнює 0,2
16	Згладжене середнє значення за модулем *NTD*	40	Кількість оцінок *NTD*, яка дорівнює 0,3
17	Згладжене середнє абсолютне значення *NTD*	41	Кількість оцінок *NTD*, яка дорівнює 0,4
18	Сума максимального і мінімального значень оцінок *NTD*	42	Кількість оцінок *NTD*, яка дорівнює 0,5
19	Різниця між максимальним і мінімальним значеннями оцінок *NTD*	43	Кількість оцінок *NTD*, яка дорівнює 0,6
20	Максимальне значення оцінки *NTD*	44	Кількість оцінок *NTD*, яка дорівнює 0,7
21	Мінімальне значення оцінки *NTD*	45	Кількість оцінок *NTD*, яка дорівнює 0,8
22	Сума всіх оцінок *NTD*	46	Кількість оцінок *NTD*, яка дорівнює 0,9
23	Середнє значення знаків оцінок *NTD*	47	Кількість оцінок *NTD*, яка дорівнює 1
24	Стандартне відхилення знаків оцінок *NTD*	48	Кількість оцінок *NTD* більше 1

Для визначення темпоральної спрямованості з цільовою ознакою, яка набуває значення (–1, 1) або (–1, 0,1), для кожної пари документів можна використати різні класифікаційні інтелектуальні моделі машинного навчання.

Використовуючи датасет датованих новин, можна створити навчальний набір даних для моделей штучного інтелекту, попарно порівнюючи новини визначеним вище способом з метою передбачення темпоральної спрямованості на рівні документів. В результаті, ці моделі передбачатимуть імовірність того, що один документ хронологічно передує іншому, формуючи таким чином відносний порядок подій між різними документами.

Практичне застосування технології

Співавтором цієї статті Б. С. Білецьким створено датасет на основі україномовних новин із соцмереж. Перевагою цієї платформи є обмежена кількість слів у новинах (в середньому 40…50

слів). Оскільки новини публікуються майже одночасно з подією, у тексті публікації не вказуються часові обставини, що робить ці новини особливо корисними для нашого дослідження. Зібрано датасет, який містить 127 000 новин.

За допомогою інструментів оброблення природної мови у новинах здійснено очищення тексту від неукраїнських символів, посилань, стоп-слів і стоп-фраз, видалення пунктуації, лематизацію слів. Створено словник унікальних слів, який нараховує 1654 лексичні одиниці.

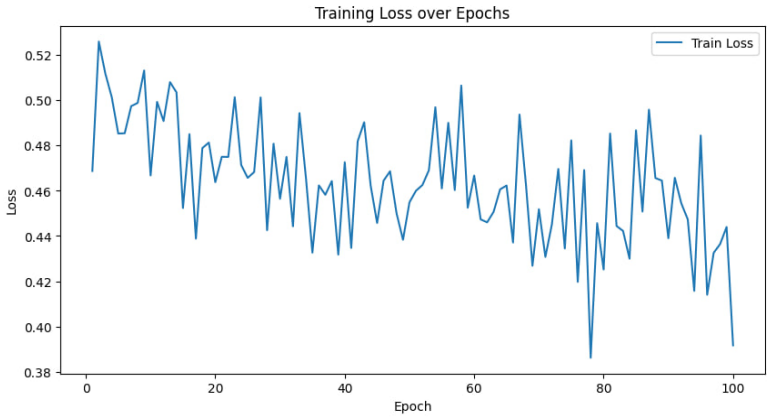

Рис. 1. Графік залежності функції втрат від епох навчання за метрикою MSE

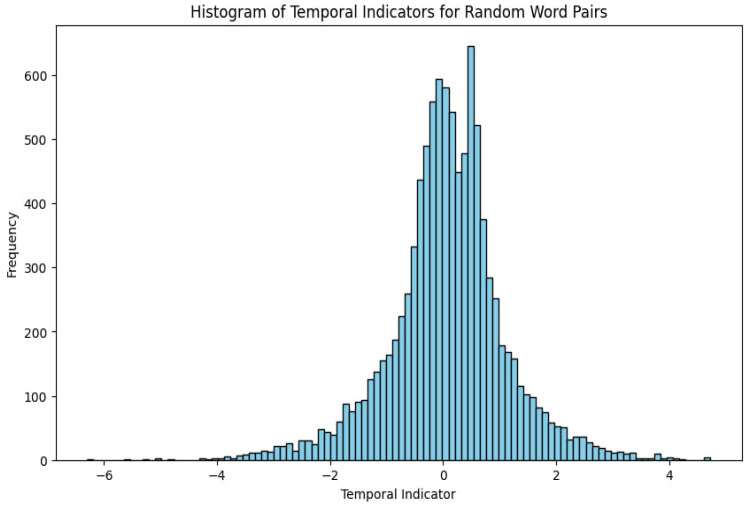

Рис. 2. Гістограма передбачень для 1000 випадкових пар слів зі словника

Для обчислення темпоральної спрямованості пар слів документів використано нейронну мережу з 5 прихованими шарами. Функція активації — ReLU, Dropout = 0,2, метрика — MSE («Mean Square Error» — з англ.: «Середньоквадратична похибка» [5]). На рис. 1 показано графік залежності функції втрат (Loss) за цією метрикою від епох навчання. А на рис. 2 показана гістограма передбачень для 1000 випадкових пар слів зі словника, побудованого на основі цього датасету.

Використовуючи модель передбачення значень показника темпоральної спрямованості на рівні слів, вибрано 400 текстів датованих новин, які порівнювалися між собою. Для кожної пари створено матриці оцінок, агреговані за 48 ознаками з табл. 1. Для кожної пари новин визначався цільовий показник: 1 — для пари новин, що знаходяться в прямому хронологічному порядку, та –1 — для пари новин, розташованих у зворотному хронологічному порядку.

Усі дані розбито на навчальний (80 %) і валідаційний (20 %) датасети. Як зазвичай, на навчальному датасеті навчено класифікаційні моделі штучного інтелекту і потім за ними передбачено класи серед даних валідаційного датасету. У табл. 2 подано отримані відповідні значення метрики «Accuracy_score» (відсоток правильно передбачених класів [5]) для навчального («Accuracy_training») та валідаційного («Accuracy_validation») датасетів окремо (табл. 2).

Таблиця 2

Результати навчання моделей штучного інтелекту

№	Назва моделі	Accuracy_training	Accuracy_validation
1	Gradient Boosting Classifier	0,954534	0,897601
2	Random Forest Classifier	0,948678	0,748106
3	Decision Tree	0,874904	0,689773
4	Logistic Regression	0,750984	0,585606

Результати показали, що модель «Gradient Boosting Classifier» бібліотеки Sklearn досягла найвищого показника точності Accuracy_validation — 89,76 %, при цьому відхилення від Accuracy_training не перевищує 10 %, що є доволі гарним результатом.

Висновки

Запропоновано підхід до визначення темпоральної спрямованості в текстах новин, які не містять явних часових маркерів. Введено новий термін «темпоральна спрямованість», який відображає ймовірнісний підхід до оцінювання порядку появи слів на основі статистичних закономірностей у текстових даних. Запропонований підхід використовує глибокі нейронні мережі для моделювання ймовірності того, що одне слово з'являється раніше за інше, шляхом попарного порівняння слів у новинах.

Практичне застосування підходу продемонстровано на датасеті з 127 000 новин, зібраних із соцмереж. Результати показали високу точність у хронологічному впорядкуванні документів. У майбутньому планується розширити метод дослідження на тексти інших жанрів та мов, а також удосконалити модель для роботи з більшими обсягами даних і складнішими мовними конструкціями.

Протестовано декілька інтелектуальних моделей для класифікації темпоральної спрямованості на рівні документів. Найкращий результат 89,76 % за метрикою «Accuracy_score» на валідаційному датасеті показала модель «Gradient Boosting Classifier».

Розроблена технологія може бути застосована для автоматичного відтворення хронології новинних подій, навіть коли явних часових маркерів немає. Це може бути корисно для автоматизації аналізу великих масивів новин та встановлення зв'язків між подіями у різних контекстах.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] W. Xiang, and B. Wang, "A survey of event extraction from text," *IEEE Access*. vol. 7, pp. 173111-173137, 2019.

[2] S. Zhang, L. Huang, and Q. Ning, "Extracting Temporal Event Relation with Syntactic-Guided Temporal Graph Transformer," *arXiv*: 2104, 09570, 2021.

[3] X. Xu, T. Gao, Y. Wang, and X. Xuan, "Event temporal relation extraction with attention mechanism and graph neural network," *Tsinghua Sci. Technol.*, vol. 27, pp. 79-90, 2021.

[4] M. Ballesteros, O. Papadopoulou, and N. Goyal, "Severing the edge between before and after: Neural architectures for temporal ordering of events," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 5068-5079.

[5] Q. Ning, Z. Feng, and D. Roth, "A Structured Learning Approach to Temporal Relation Extraction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, 2017, pp. 1027-1037.

[6] T. Goyal, and G. Durrett, "Embedding time expressions for deep temporal ordering models," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4401-4411, 2019.

[7] Y. Liu, J. Ma, and P. Li, "Predicting higher-order patterns in temporal networks," in *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, 2021, pp. 3219-3228.

[8] W. Xia, Y. Li, and S. Li, "Graph neural point process for temporal interaction prediction," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2023, pp. 1-10.

[9] Q. Ning, S. Subramanian, and D. Roth, "An Improved Neural Baseline for Temporal Relation Extraction," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6203-6209.

[10] A. Naik, L. Breitfeller, and C. Rose, "TDDiscourse: A dataset for discourse-level temporal ordering of events," *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 2019, pp. 239-249.

[11] В. Б. Мокін, і М. В. Дратований, *Наука про дані: машинне навчання та інтелектуальний аналіз даних*, електр. навч. посіб. комбінованого (локального та мережевого) використання. Вінниця, Україна: ВНТУ, 2024, 258 с. [Електронний ресурс]. Режим доступу: https://docs.vntu.edu.ua/card.php?id=8163.

Рекомендована кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 11.12.2024

Білецький Богдан Сергійович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: bohdanbeletskyi@gmail.com ;

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій, e-mail: vbmokin@gmail.com

B. S. Biletskyi[1]
V. B. Mokin[1]

Determining of Temporal Directionality in Texts: a Neural Network-Based Approach for Chronological Ordering Based on Pairwise Word Analysis

The article presents a neural network approach for determining the temporal orientation in texts, which allows reconstructing the chronology of events even in the absence of explicit time markers. This approach determines the probabilistic order of words in texts, taking into account their statistical and linguistic relationships. In contrast to traditional approaches that rely on explicit temporal expressions or publication dates, the proposed approach allows to estimate the order of events based on the identified relationships between pairs of words in documents, describing events.

To analyze the temporal orientation, neural networks are used to model the relationships between words by comparing their occurrence in texts in pairs. Formulas have been developed to calculate temporal orientation indicators based on the frequency of occurrence of words in dated texts. The obtained indicators are normalized, this provides a better interpretation of the results.

Based on these indicators, a set of features was formed to train machine learning models according to various criteria. To test the effectiveness, we created a Ukrainian-language corpus of 127,000 social media news and applied several models: Gradient Boosting Classifier, Random Forest Classifier, Decision Tree, and Logistic Regression. As an example, 48 features that characterize the news, were selected. The experiments revealed that the Gradient Boosting Classifier model showed the best result with an accuracy of 89.76 % on the validation dataset, which exceeded the accuracy of other models such as Random Forest (74.81%) and Decision Tree (68.97 %).

The proposed approach proved to be effective in modeling the chronological relationships between events, which is important for text automation tasks. The approach can be used to analyze news, chronologically organize historical events, and work with text data in large arrays.

Keywords: intelligent technologies, machine learning, artificial intelligence, neural networks, natural language processing, temporal directionality, information technology.

Biletskyi Bohdan S. — Post-Graduate Student of the Chair of System Analysis and Information Technologies, e-mail: bohdanbeletskyi@gmail.com ;

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technologies, e-mail: vbmokin@vntu.edu.ua