УДК 004.855.5
DOI https://doi.org/10.32689/maup.it.2022.3.6

Роман ШАПТАЛА
аспірант кафедри системного проектування, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», корпус 14, вул. Політехнічна 14-б, 03056, Київ, Україна (r.shaptala@gmail.com)

ORCID: 0000-0002-4367-5775

Геннадій КИСЕЛЬОВ
кандидат технічних наук, доцент, заступник завідувача кафедри системного проектування, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», корпус 14, вул. Політехнічна 14-б, 03056, Київ, Україна (g.kyselov@gmail.com)

ORCID: 0000-0003-2682-3593

Roman SHAPTALA
PhD student at the Department of System Design, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", build. 14, Politekhnichna 14-b str., Kyiv, Ukraine, postal code 03056 (r.shaptala@gmail.com)

Gennadiy KYSELOV
Candidate of technical sciences, Associate professor, Deputy head of the Department of System Design, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", build. 14, Politekhnichna 14-b str., Kyiv, Ukraine, postal code 03056 (g.kyselov@gmail.com)

Бібліографічний опис статті: Шаптала, Р., Кисельов, Г. (2022). Класифікація текстових документів з використанням доповнення векторних представлень документів графовими представленнями елементів словника синонімів. *Інформаційні технології та суспільство*, 3 (5), 49–55. DOI:

Bibliographic description of the article: Shaptala, R., Kyselov, G. (2022). Klasyfikaciya tekstovyh dokumentiv z vykorystannyam dopovnennya vektornyh predstavlen dokumentiv grafovymy predstavlennyamy elementiv slovnyka synonimiv [Document classification via augmentation of document embeddings with graph embeddings of synonyms dictionary]. *Informatsiini tekhnolohii ta suspilstvo – Information technology and society*, 3 (5), 49–55. DOI:

КЛАСИФІКАЦІЯ ТЕКСТОВИХ ДОКУМЕНТІВ З ВИКОРИСТАННЯМ ДОПОВНЕННЯ ВЕКТОРНИХ ПРЕДСТАВЛЕНЬ ДОКУМЕНТІВ ГРАФОВИМИ ПРЕДСТАВЛЕННЯМИ ЕЛЕМЕНТІВ СЛОВНИКА СИНОНІМІВ

*Стаття присвячена оцінці впливу методів доповнення графовими представленнями елементів словника синонімів векторних представлень документів на якість класифікації даних документів у малоресурсному середовищі. Дослідження таких середовищ є актуальним завданням, адже більшість мов світу, а також вузькоспеціалізовані прикладні області підпадають під даний критерій – даних для побудови та тренування сучасних потужних моделей машинного навчання не достатньо. **Метою роботи** є покращення якості класифікації документів у малоресурсному середовищі за допомогою доповнення їх інформацією зі словника синонімів через його кодування. Дослідження виконано через аналіз та використання сучасних напрацювань у області математичного моделювання, машинного навчання, обробки природних мов та науки про дані.*

***Наукова новизна роботи** полягає у тому, що пропонується векторна модель слів зі словника синонімів, яка на відміну від інших працює на основі представлень окремих вузлів графу словника, а отже може бути використана і в інших задачах обробки текстових даних. У цьому може допомогти трансферне навчання – підхід, що дозволяє комбінувати щільні векторні представлення у нейромережевих методах. При цьому вибір методу побудови векторних представлень словника синонімів напряму впливає на якість результатів, а також швидкодію та вимоги до апаратного забезпечення при їх використанні. Також у роботі представлено набір кроків передобробки та спосіб перетворення словника у граф для моделювання. Як **висновок**, у статті показано, що запропонований метод здатен збільшити F1-міру точності класифікації документів у малоресурсному середовищі на 2-3% на прикладі класифікації петицій до Київської міської ради за темами. Найвищий приріст якості було отримано за допомогою методу побудови векторних представлень графу Node2Vec, що працює на основі випадкових блукань, та не вимагає великої кількості ресурсів для навчання.*

***Ключові слова:** обробка природної мови, векторні представлення, класифікація документів, математична модель, машинне навчання, нейронні мережі, малоресурсне середовище.*

DOCUMENT CLASSIFICATION VIA AUGMENTATION OF DOCUMENT EMBEDDINGS WITH GRAPH EMBEDDINGS OF SYNONYMS DICTIONARY

The article is devoted to the assessment of the influence of the methods of augmentation of vector representations of documents with graph representations of the elements of the synonym dictionary on the quality of the classification of these documents in a low-resource environment. The study of such environments is an important task, because most of the world's languages, as well as highly specialized application areas, fall under this criterion – there is not enough data for building and training modern powerful machine learning models. The ***main goal*** *of this article is to improve the quality of document classification in a low-resource environment by augmenting them with information from the dictionary of synonyms through the encoding of the latter. The research was carried out through the analysis and use of modern developments in the field of mathematical modeling, machine learning, natural language processing and data science.*

The ***scientific novelty*** *of the work lies in the fact that a vector model of words from the dictionary of synonyms is proposed, which, unlike others, works on the basis of representations of individual nodes of the dictionary graph, and therefore can be used in other text data processing tasks. This can be helped by transfer learning, an approach that allows combining dense vector representations in neural network methods. At the same time, the choice of the method of building vector representations of the dictionary of synonyms directly affects the quality of the results, as well as the speed and requirements for hardware when using them. Also, the work presents a set of preprocessing steps and a method of converting a dictionary into a graph for modeling. As a* ***conclusion****, the article shows that the proposed model is able to increase the F1-score of document classification in a low-resource environment by 2-3% using the example of the classification of petitions to the Kyiv City Council by topic. The highest quality gain was obtained using the Node2Vec method of constructing graph vector representations, which works on the basis of random walks and does not require a large amount of training resources.*

Key words: *natural language processing, vector embeddings, document classification, mathematical model, machine learning, neural networks, low-resource.*

Актуальність. Попри стрімкий розвиток технологій та методів штучного інтелекту у контексті обробки природних мов, більшість уваги зосереджена на обмеженій кількості високоресурсних мов та прикладних областей, таких як англійська мова чи сентимент аналіз текстів у соціальних мережах, де наявна велика кількість джерел даних для побудови статистичних моделей. У випадку, коли ж даних – обмаль, багато припущень, на яких базуються популярні підходи не виконуються, а отже у малоресурсних середовищах потрібно шукати нові методи для успішного вирішення завдань. Для цього зазвичай вдаються до розробки підходів, які б могли використати сторонню інформацію аби отримати з неї певне додаткове «розуміння» контексту та середовища для вирішення оригінального завдання. У статті пропонується використати словник синонімів для того, щоб збільшити якість класифікації документів при обробці природних мов у малоресурсному середовищі. Тож **метою дослідження** є перевірка гіпотези, що кодування словника синонімів за допомогою векторних представлень здатне покращити результати роботи інших моделей машинного навчання.

Аналіз останніх досліджень та публікацій. Методи обробки природних мов у малоресурсному середовищі можна поділити на кілька груп [1]: створення додаткових розмічених даних та трансферне навчання. Перша група включає в себе підгрупи: доповнення даних [2], міжмовні проекції [3] та віддалений нагляд [4]. Друга група ділиться на: методи векторних представлень [5], багатомовні моделі мов [6] та адаптацію домену моделі мови [7]. Попри доведену ефективність даних підходів у певних ситуаціях, кожна з підгруп спирається на припущення існування даних або моделей, які мають певну інформацію про середовище, наприклад при використанні методу міжмовних проекцій очікується наявність паралельних корпусів між малоресурсною та багаторесурсними мовами. Наше дослідження пропонує використати словник синонімів, який наявний у багатьох мовах як джерело додаткової інформації для моделювання текстових даних.

Виклад основного матеріалу дослідження. Яким чином використати інформацію зі словника синонімів? Дослідники [8] пропонують робити доповнення даних за допомогою генерації нових екземплярів у тренувальній вибірці, у яких замінені слова-синоніми між собою. Такий підхід має ряд недоліків: розподіли слів у вибірці перестають бути репрезентативними, додається шум у набір для тренування, вибір кількості додатково згенерованих даних напряму впливає на результат, а також даний підхід важко перевикористати для інших завдань – правила заміни слів пишуться залежно від прикладної області. Для вирішення даних проблем пропонується моделювати словник синонімів як граф, де окремі вузли – це слова, а ребра між ними позначають синонімію. Після цього за допомогою методу побудови векторних представлень вузлів графу кожному слову знайти відповідний вектор. При цьому, слова, що знаходяться у графі близько повинні мати близькі за косинусною відстанню представлення. Запропоновану модель на високому рівні схематично зображено на рис. 1.

Як обрати метод побудови векторних представлень вузлів? За принципом дії їх поділяють на декілька груп [9]: на основі факторизації, на основі випадкових блукань та на основі глибокого навчання. Так, методи на основі факторизації розкладають матрицю, що репрезентує граф, на компоненти з бажани-

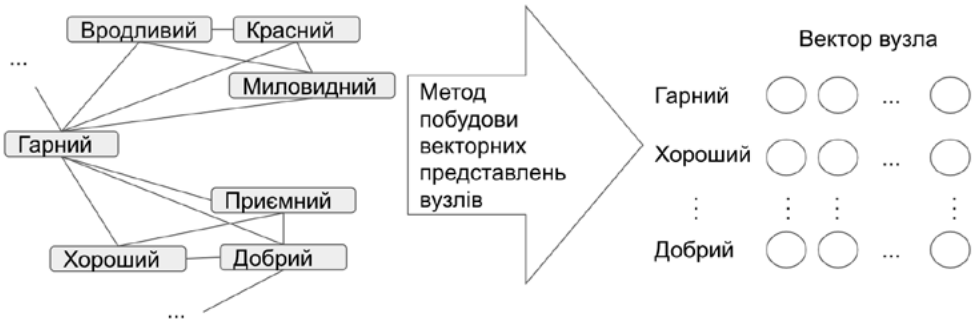

Рис. 1. Схема моделі словника синонімів

ми властивостями. Такою матрицею зазвичай виступає матриця суміжності вузлів, матриця Лапласа чи матриця ймовірностей переходу з одного вузла графу в інший. Прикладами даних методів є факторизація графу [10], HOPE [11], LLE [12], GraRep [13], Лапласівські проекції [14]. Перевагою цієї групи методів є простота та маленька кількість додаткових гіперпараметрів, але є і суттєві недоліки – для великих графів відповідні матриці вимагають забагато пам'яті – як мінімум квадратичну функцію від кількості вузлів, а отже обмежені у можливих застосуваннях. Метод факторизації графу працює має часову складність пропорційну кількості ребер і розкладає матрицю суміжності графу на множники. На відміну від нього, HOPE намагається явно закодувати подібності вищих порядків через мінімізацію схожості між елементами графу. У свою чергу, метод LLE робить припущення, що кожне представлення вузла є лінійною комбінацією представлень його сусідів. Унікальність методу GraRep полягає у використанні матриці ймовірностей переходів між вузлами та тому, що його часова складність не залежить від кількості ребер у графі, але кубічна пропорційно кількості вузлів. Метод Лапласівських проекцій розкладає Лапласіаном графа на власні вектори, отримуючи таким чином векторні представлення вузлів графа.

Методи на основі випадкових блукань вирішують проблему з пам'яттю наявну у попередньої групи методів, адже є ітеративними та працюють з графом напряму, а не з його матричним представленням. Такі методи працюють у два етапи – генерація випадкових блукань, коли створюються послідовності вузлів при випадковому обході графу, та навчання представлень, коли на основі послідовностей налаштовуються вектори для моделювання зв'язків різних порядків. Недоліком є велика кількість гіперпараметрів, що потрібно перебирати, таких як параметри випадкового блукання, кількість згенерованих послідовностей, кількість ітерацій при тренуванні. Популярні реалізації таких методів: Node2Vec [15], Walklets [16] та DeepWalk [17]. DeepWalk є першим онлайн алгоритмом навчання векторних представлень на графах, а тому легко масштабується. Саме його автори вперше почали генерувати випадкові блукання, а потім кодувати векторні представлення так, щоб з їх допомогою можна було прогнозувати наступний вузол у послідовності. Node2Vec узагальнив цей процес додавши можливість керувати генерацією випадкових блукань, а саме фокусом на обхід графу в ширину або в глибину. За це відповідають гіперпараметри p та q відповідно. Таким чином можна сказати, що DeepWalk є частковим випадком застосування Node2Vec при p=q=1. Метод Walklets додає своєрідну особливість до даного ітеративного процесу – він дозволяє при генерації пропускати деяких сусідів, що може спонукати модель краще відтворювати зв'язки вищих порядків.

Останнім часом популярності набула третя група методів, яка працює на основі глибокого навчання. Основна ідея таких методів – за допомогою глибоких автокодерів зменшити розмірність матричних представлень графів. При цьому конкретні реалізації даних методів відрізняються типами автокодерів та функціями втрат, що оптимізуються під час навчання. Такі методи вимагають менше пам'яті ніж перша група і при цьому здатні працювати ітеративно як друга група, але є повільнішими, тому що кількість параметрів при оптимізації значно вища. До даної групи відносяться підходи SDNE [18], DNGR [19] та GCN [20]. Підхід SDNE намагається одночасно оптимізувати збереження близькостей між предствленнями вузлів першого та другого порядку за допомогою глибоких автокодерів. У свою чергу DNGR використовує шумопоглинаючі автокодери для розкладу матриці ймовірностей переходів побудованої на основі тих же випадкових блукань. На жаль, обидва алгоритма можуть бути обчислювально складними, адже на кожному кроці приймають на вхід увесь граф. Цю проблему вирішують GCN – графові згорткові мережі, що визначають операцію згортки на графі, таким чином агрегуючи локальну структуру графу на нижніх шарах, та глобальну на верхніх.

Для перевірки впливу доповнення даних інформацією зі словника синонімів закодованого методами побудови графових представлень було розроблено наступний експеримент. Набір даних – петиції до

Київської міської ради (представлений у [21]), завдання – класифікація документів за темами, набір поділений на вибірки для тренування (75%) та тестування (25%). Малоресурсність середовища забезпечена двома аспектами набору даних: документи у ньому написані українською мовою, вузькоспеціальною прикладною областю урбаністики. Базова модель без доповнення даних – усереднені закодовані Word2Vec підходом вектори слів петицій як ознаки, що подаються на вхід багатошаровому перцептрону як класифікатору. В усіх експериментах кожна частина загального процесу включає підбір гіперпараметрів за допомогою сіткового пошуку. Таким чином випадковість ініціалізації не впливатиме на загальний результат. Метрика якості класифікації – зважена F1-міра, обрана через незбалансованість набору даних за класами, що є частим ускладненням при роботі у малоресурсному середовищі. Точність у такому випадку показувала б надто оптимістичні результати, які не виражали б істинну якість моделі.

Для побудови графу було викачано україномовний словник синонімів з Офіційного сайту української мови [22]. Але перед його кодуванням важливо провести передобробку цих даних. Вона включає такі кроки: видалення зайвої для експерименту інформації такої як приклади вжитку та скорочень; дедублікація пар синонімів, адже пропонується будувати ненапрямлений граф, а у словнику є зв'язки між словами у обидві сторони; так як власні назви у словнику не зустрічаються, усі слова приведені до нижнього регістру для легшого пошуку відповідників у текстах петицій. Таким чином на вхід процесу побудови графу надається список чистих пар «синонім-синонім». При побудові, усі ці пари з'єднуються між собою ненапрямленими ребрами, утворюючи кластери синонімічних зв'язків. Результуючий граф має 44664 вузли та 391047 ребер, а отже є достатньо великим для врахування обмежень за пам'яттю та швидкодією при виборі методів побудови векторних представлень вузлів. Середня степінь вершини у графі складає 8.755, що свідчить про високу зв'язність графу, а також, що його локальна структура відіграватиме важливу роль при моделюванні.

Яким чином поєднати вектори петицій та вектори слів зі словника синонімів? Для цього можна скористатись простими методами з області мультимодального навчання, а саме методами конкатенації та зваженої суми векторних представлень. У першому випадку розмірність результуючого вектора, який піде на вхід моделі класифікації, являється сумою розмірностей вектора петиції та вектора вузла графу словника синонімів, тоді ж як у другому випадку накладається обмеження, що вищевказані вектори повинні бути однієї розмірності, відповідно при виконанні операції зваженої суми, результуючий вектор теж матиме такий розмір. Більш складні методи злиття векторних представлень такі як методи на основі механізму уваги вимагають підбір додаткового набору параметрів, що ускладнює оптимізацію, а також вимагає більше даних для навчання. У малоресурсних середовищах застосування таких методів не практичне, тому у дослідженні перевіряються лише методи злиття на основі простих операцій.

Подібно процедурі усереднення векторів слів Word2Vec, яка є базовою моделлю у дослідженні, варто визначити процедуру агрегації окремих векторів слів зі словника синонімів. Ми пропонуємо також усереднювати дані вектори, ігноруючи ті слова з документу, що не існують у словнику. Саме на цьому етапі важливо максимізувати кількість слів, що перетинаються між множиною слів з документів та множиною слів зі словника. Тому передобробка обох джерел даних має суттєвий вплив на результат. Також, враховуючи, що слова у словниках зазвичай подаються у називному відмінку або інфінітиві, а у документах залежно від контексту можуть зустрічатись у інших формах, процедура їх співставлення не очевидна. Ідеальною процедурою у даному випадку було б використання двох додаткових моделей – моделі визначення частини мови та моделі побудови словоформ. Малоресурсність середовища гарантує, що таких моделей високої якості не існує, а даних для їх розробки обмаль. Тому для процедури співставлення пропонується використати алгоритми пошуку найближчого слова за міжрядковою відстанню. Це додає шум в фінальні представлення, які йдуть на вхід класифікатору, але забезпечує вище покриття слів при співставленні документів та словника синонімів.

Експериментальні результати. Для підтвердження гіпотези, що злиття векторних репрезентацій документів та елементів словника синонімів було проведено набір експериментів у різних комбінаціях способів реалізації запропонованого методу. Таким чином порівнювались базова модель з методами її покращення за допомогою інформації зі словника синонімів: доповнення даних через заміну синонімів та злиття з векторним представленням графа синонімів. При цьому останній залежить від вибору методу векторного представлення словника синонімів (побудови графового представлення), а також методу злиття векторних представлень (конкатенація чи зважена сума). Серед методів векторного представлення словника синонімів було порівняно методи факторизації графу, HOPE, LLE, Лапласівські проекції, GraRep, Node2Vec, Walklets та GCN. Через розмір графу, а також пропорційну кількості ребер часову складність алгоритмів HOPE, LLE та Лапласівські проекції виявились не практичними у застосуванні в даному середовищі. Кубічна складність у кількості вузлів алгоритму GraRep теж завадила йому завершитись успішно, тому ці алгоритми не рекомендуються у застосуванні при роботі зі

словником синонімів. Так як застосування методу Node2Vec з гіперпраметрами q та p рівними одиниці рівнозначне застосуванню методу DeepWalk, а саме дані значення виявились оптимальними, у результатах зазначається лише перший. Серед методів на основі глибокого навчання для експериментів було обрано GCN, адже він є узагальненням SDNE та DNGR, а отже покриває їх результати при правильному підборі гіперпараметрів. Значення результуючих метрик для порівняння наводиться у таблиці 1.

Таблиця 1

Порівняння F1-міри запропонованих підходів

Метод обробки природної мови у малоресурсному середовищі	**Метод векторного представлення словника синонімів**	**Метод злиття векторних представлень**	**Зважена F1-міра**
Базова модель	-	-	0.629
Доповнення даних через заміну синонімів	-	-	0.631
Злиття з векторним представленням графа синонімів	Факторизація графу	Конкатенація	0.639
Злиття з векторним представленням графа синонімів	Факторизація графу	Зважена сума	0.650
Злиття з векторним представленням графа синонімів	Node2Vec	Конкатенація	0.640
Злиття з векторним представленням графа синонімів	Node2Vec	Зважена сума	**0.659**
Злиття з векторним представленням графа синонімів	Walklets	Конкатенація	0.638
Злиття з векторним представленням графа синонімів	Walklets	Зважена сума	0.657
Злиття з векторним представленням графа синонімів	GCN	Конкатенація	0.645
Злиття з векторним представленням графа синонімів	GCN	Зважена сума	0.657

Як бачимо, найвищий приріст метрики – 3% порівняно з базовою моделлю надає злиття методом зваженої суми з векторним представленням графа синонімів побудованим за алгоритмом Node2Vec. Для перевірки статистичної значущості результату було проведено Хі-квадрат тест за методом МакНемара [23], а саме було сформульовано гіпотезу, що результати класифікації базовим методом та запропонованим методом – рівні. Вхідні дані для проведення тесту відображені у таблиці 2. Рівень статистичної значущості приймаємо як 0.05. Фінальне значення статистики вищезазначеного тесту рівне 4.056, що відповідає p-значенню у 0.044, що менше прийнятого рівня статистичної значущості 0.05, а отже початкова гіпотеза про рівність результатів класифікації базовим та запропонованим методами відхиляється.

Таблиця 2

Таблиця невідповідностей для Хі-квадрат тесту за методом МакНемара

		Класифікація запропонованим методом	
		правильна	**помилкова**
Класифікація базовим методом	правильна	757	59
	помилкова	83	350

За допомогою трансферного навчання отримані векторні представлення словника синонімів можна перевикористати для покращення якості вирішення інших завдань. Для цього варто лише пересвідчитись, що кроки передобробки, описані раніше, не прибирають корисної інформації для нового завдання, а отже можна через обраний метод злиття додавати існуючі вектори до іншої класифікаційної архітектури. Якщо ж описана передобробка для нового завдання не оптимальна, варто змінити дані кроки, застосувати нову процедуру на словнику синонімів та перетренувати метод побудови векторних представлень графу на оновлених даних.

Висновки. У статті запропоновано модель словника синонімів української мови, а також метод, який поєднує класичний підхід до класифікації документів з векторними представленнями словника синонімів, що дозволяє збільшити зважену F1-міру класифікації документів на 3% при обробці природної мови у малоресурсних середовищах. Значущість отриманих результатів була підтверджена за допомогою Хі-квадрат тесту за методом МакНемара з рівнем статистичної значущості 0.05. Серед методів побудови векторних представлень графу в даному середовищі найкращі результати показав Node2Vec, а серед методів злиття векторних представлень рекомендується обрати метод зваженої суми. Роботу можна розвивати для дослідження та моделювання інших типів словників за допомогою графових векторних представлень, створення та модифікації методів побудови графових векторних представлень, які були б оптимізовані під структуру графів на основі словників, а також перевірку ефективності дії методу в інших малоресурсних середовищах.

Список використаних джерел:

1. Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2545–2568).

2. Perez, L., & Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. Retrieved from https://arxiv.org/abs/1712.04621.

3. Eskander, R., Muresan, S., & Collins, M. (2020). Unsupervised Cross-Lingual Part-of-Speech Tagging for Truly Low-Resource Scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4820–4831). Association for Computational Linguistics (ACL). https://doi.org/10.18653/V1/2020.EMNLP-MAIN.391.

4. Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 1003–1011). Association for Computational Linguistics. Retrieved from https://aclanthology.org/P09-1113.

5. Collobert, R., Weston, J., Com, J., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, *12*, 2493–2537. https://doi.org/10.5555/1953048.2078186.

6. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *International Conference on Machine Learning. PMLR*.

7. Gururangan, S., Marasovi´cmarasovi´c, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8342–8360). Association for Computational Linguistics. Retrieved from https://github.com/allenai.

8. Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 6382–6388). Retrieved from http://github.

9. Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, *151*, 78–94. https://doi.org/10.1016/J.KNOSYS.2018.03.022

10. Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V., & Smola, A. J. (2013). Distributed Large-Scale Natural Graph Factorization. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 37–48). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2488388.2488393.

11. Ou, M., Cui, P., Pei, J., Zhang, Z., & Zhu, W. (2016). Asymmetric Transitivity Preserving Graph Embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1105–1114). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939751.

12. T., R. S., & K., S. L. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, *290*(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323.

13. Cao, S., Lu, W., & Xu, Q. (2015). GraRep: Learning Graph Representations with Global Structural Information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 891–900). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2806416.2806512.

14. Belkin, M., & Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, *15*(6), 1373–1396. https://doi.org/10.1162/089976603321780317.

15. Grover, A., & Leskovec, J. (2016). Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 855–864). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939754.

16. Perozzi, B., Kulkarni, V., Chen, H., & Skiena, S. (2017). Don't Walk, Skip! Online Learning of Multi-Scale Network Embeddings. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 258–265). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3110025.3110086.

17. Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 701–710). https://doi.org/10.1145/2623330.2623732.

18. Wang, D., Cui, P., & Zhu, W. (2016). Structural Deep Network Embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1225–1234). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939753.

19. Cao, S., Lu, W., & Xu, Q. (2016). Deep Neural Networks for Learning Graph Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1 SE-Technical Papers: Machine Learning Applications). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/10179.

20. Kipf, T. N., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017 – Conference Track Proceedings*. International Conference on Learning Representations, ICLR. Retrieved from https://arxiv.org/abs/1609.02907v4.

21. Samvelyan, A., Shaptala, R., & Kyselov, G. (2020). Exploratory data analysis of Kyiv city petitions. In *2020 IEEE 2nd International Conference on System Analysis Intelligent Computing (SAIC)* (pp. 1–4). https://doi.org/10.1109/SAIC51296.2020.9239185.

22. Офіційний сайт Української мови. URL: https://ukrainskamova.com.

23. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996.

References:

1. Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2545–2568).

2. Perez, L., & Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. Retrieved from https://arxiv.org/abs/1712.04621.

3. Eskander, R., Muresan, S., & Collins, M. (2020). Unsupervised Cross-Lingual Part-of-Speech Tagging for Truly Low-Resource Scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4820–4831). Association for Computational Linguistics (ACL). https://doi.org/10.18653/V1/2020.EMNLP-MAIN.391.

4. Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 1003–1011). Association for Computational Linguistics. Retrieved from https://aclanthology.org/P09-1113.

5. Collobert, R., Weston, J., Com, J., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, 12, 2493–2537. https://doi.org/10.5555/1953048.2078186.

6. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. In *International Conference on Machine Learning. PMLR*.

7. Gururangan, S., Marasovi´cmarasovi´c, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8342–8360). Association for Computational Linguistics. Retrieved from https://github.com/allenai.

8. Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 6382–6388). Retrieved from http://github.

9. Goyal, P., & Ferrara, E. (2018). Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151, 78–94. https://doi.org/10.1016/J.KNOSYS.2018.03.022.

10. Ahmed, A., Shervashidze, N., Narayanamurthy, S., Josifovski, V., & Smola, A. J. (2013). Distributed Large-Scale Natural Graph Factorization. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 37–48). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2488388.2488393.

11. Ou, M., Cui, P., Pei, J., Zhang, Z., & Zhu, W. (2016). Asymmetric Transitivity Preserving Graph Embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1105–1114). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939751.

12. T., R. S., & K., S. L. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323.

13. Cao, S., Lu, W., & Xu, Q. (2015). GraRep: Learning Graph Representations with Global Structural Information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 891–900). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2806416.2806512.

14. Belkin, M., & Niyogi, P. (2003). Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, *15*(6), 1373–1396. https://doi.org/10.1162/089976603321780317.

15. Grover, A., & Leskovec, J. (2016). Node2vec: Scalable Feature Learning for Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 855–864). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939754.

16. Perozzi, B., Kulkarni, V., Chen, H., & Skiena, S. (2017). Don't Walk, Skip! Online Learning of Multi-Scale Network Embeddings. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 258–265). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3110025.3110086.

17. Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). DeepWalk: Online learning of social representations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 701–710). https://doi.org/10.1145/2623330.2623732.

18. Wang, D., Cui, P., & Zhu, W. (2016). Structural Deep Network Embedding. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1225–1234). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939753.

19. Cao, S., Lu, W., & Xu, Q. (2016). Deep Neural Networks for Learning Graph Representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30 (1 SE-Technical Papers: Machine Learning Applications). Retrieved from https://ojs.aaai.org/index.php/AAAI/article/view/10179.

20. Kipf, T. N., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017 – Conference Track Proceedings*. International Conference on Learning Representations, ICLR. Retrieved from https://arxiv.org/abs/1609.02907v4.

21. Samvelyan, A., Shaptala, R., & Kyselov, G. (2020). Exploratory data analysis of Kyiv city petitions. In *2020 IEEE 2nd International Conference on System Analysis Intelligent Computing (SAIC)* (pp. 1–4). https://doi.org/10.1109/SAIC51296.2020.9239185.

22. Oficiinyi sait Ukrainskoi movy [Official website of Ukrainian language]. Retrieved from: https://ukrainskamova.com [in Ukrainian].

23. McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996.