Одеський національний університет імені І.І. Мечникова
Міністерство освіти і науки України

Кваліфікаційна наукова
праця на правах рукопису

Бочарова Майя Юріївна

УДК 004.056: 004.65

ДИСЕРТАЦІЯ

МЕТОДИ ТА ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПРЕДМЕТНО-ОРІЄНТОВАНОГО АНАЛІЗУ ПРИРОДНОМОВНИХ ТЕКСТІВ

122 – Комп’ютерні науки
12 – Інформаційні технології

Подається на здобуття наукового ступеня доктора філософії

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело _____ М.Ю.Бочарова

Науковий керівник:
Малахов Євгеній Валерійович,
доктор технічних наук, професор

Одеса – 2025

АНОТАЦІЯ

Бочарова М.Ю. Методи та інформаційна технологія предметно-орієнтованого аналізу природномовних текстів. — Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктор філософії (PhD) за спеціальністю 122 “Комп'ютерні науки”. — Одеський національний університет імені І.І. Мечникова, Одеса, 2025.

У дисертаційній роботі представлені результати проведених здобувачем досліджень, які виконують актуальне наукове завдання створення моделей та методів предметно-орієнтованого аналізу природномовних текстів, яке має істотне значення для розвитку інформаційних технологій.

У *вступі* дисертації обґрунтовано актуальність дослідження за темою предметно-орієнтованого аналізу природномовних текстів, сформульовані мета, задачі та методи дослідження, наведено наукову новизну та практичне значення отриманих результатів, зазначено особистий внесок здобувача.

У *першому розділі* роботи досліджено актуальний стан проблеми автоматизованого аналізу документів в сфері управління персоналом із застосуванням штучного інтелекту. Показано, що обробка резюме для витягнення ключової інформації, зіставлення вакансій та резюме є необхідним елементом підвищення ефективності рекрутингу і перспективним напрямком для подальшого удосконалення і розвитку.

Показано, що застосування моделей, які використовують комп’ютерний зір, не є доцільним для обробки візуально насичених документів у сфері управління людськими ресурсами.

На основі аналізу літературних джерел обґрунтована доцільність використання контексту під час моделювання подань фраз.

Висвітлена проблема обробки документів, обсяг яких перевищує ліміт моделей, які використовуються для їх обробки.

Показано, що відсутність даних щодо впливу обсягу тренувальних зразків при автоматичній їх генерації (з використанням великих мовних моделей) на якість сумаризації документів у домені управління персоналом потребує дослідження в цьому напрямку. Потребують дослідження підходи некерованого попереднього тренування з використанням структури документів, а також функції втрат, які використовуються для попереднього тренування (зважена функція втрат).

Відзначена важливість англійської мови для поширення набутих знань щодо використання штучного інтелекту у рекрутингу.

Встановлено необхідність удосконалення крос-лінгвістичної дистиляції векторних подань для підвищення ефективності рекрутингу із застосуванням штучного інтелекту.

Встановлено доцільність дослідження впливу дистиляції на показники швидкості та якості етапів технології обробки природномовних текстів щодо аналізу резюме та зіставлення з вимогами вакансій.

У *другому розділі* розроблено методи та моделі для предметно-орієнтованої обробки природномовних текстів. В тому числі представлено новий метод безпосередньої інтеграції параметрів про стильові ознаки, де додаткові дискретні ознаки векторизуються і передаються в архітектуру “Трансформер” разом із позиційними і токеновими поданнями.

Запропоновано новий метод тренування подань назв посад, що базується на використання фраз навичок, які зазначені в описі роботи. Цей метод базується на введенні спеціального токена для виділення та представлення кожної навички у поєднанні з контрастним тренуванням з метою зіставлення усередненого подання навичок та назви посади з одного опису роботи.

Запропоновано новий метод некерованого навчання моделі з використанням структури документів. На відміну від традиційного методу, в якій позитивні пари для подальшого контрастного навчання вибирають з

документу випадковим чином, запропонований метод базуються на використанні структури документу.

Запропоновано новий метод автоматичного створення датасету вакансія-резюме, який полягає у використанні структури документа і визначеного опису останньої ролі та перетворення цього запису на опис вакансій з використанням великої мовної моделі.

Запропоновано метод скорочення тексту з урахуванням структури документу та ключових фраз. Цей метод полягає у скороченні кожної секції пропорційно до її відсоткового внеску у загальну довжину резюме на основі виділення ключових фраз.

Розроблено метод некерованого попереднього тренування для сумаризації документів у сфері управління персоналом. Цей метод полягає у використанні секції “анотація” з резюме для некерованого тренування моделі сумаризації, а також у застосуванні зваженої функції втрат, яка підвищує вагомість для токенів, які формують ключові фрази.

У *третьому розділі* представлена інформаційна технологія предметно-орієнтованого аналізу природномовних текстів, яка може бути застосована за двома напрямками: вироблення рекомендацій резюме в умовах відсутності рекрутера, та інтенсифікації процесу відбору та ранжування резюме рекрутером, що дає можливість рекрутерам швидко та зручно ознайомлюватися з рекомендованими кандидатами та відфільтровувати їх.

Представлена технологія є послідовністю застосування наступних етапів: “Сегментація”, “Парсинг”, “Сумаризація”, “Векторизація”. В результаті застосування цих етапів документ перетворюється на сукупність атрибутів, анотації та векторного подання, які зберігаються у векторній базі даних.

Показано, що для оцінювання етапів технології “AI ResJobFit” необхідно обчислювати наступні показники: *F1, Recall@N, MAP, MRR, nDCG,* $Rouge_N$

У *четвертому розділі* проводиться обґрунтування ефективності та систематизація розроблених методів для обробки природномовних текстів в сфері управління людськими ресурсами.

Встановлено, що застосування безпосередньої інтеграції параметрів про стильові ознаки (без використання комп'ютерного зору) дозволяє досягти покращення якості класифікації токенів в завданні сегментації резюме та вакансій, а також витягнення ключової інформації з них.

Показано, що новий метод навчання з контекстно-орієнтованим вирівнюванням подань фраз призводить до значного покращення якості подань фраз. Емпірично визначено, що підхід "всі негативні пари" при тренуванні в умовах асиметричного датасету при використанні функції множинних негативних втрат при ранжуванні застосовувати недоцільно, бо він призводить до зниження метрик.

Показано, що використання функції втрат на основі косинусної подібності призводить до значних покращень (на 14,2% за абсолютним показником NMI) у порівнянні з використанням функції втрат середньоквадратичної похибки при дистиляції векторних подань текстів з добре натренованої моделі-вчителя.

Представлено новий і перший у своєму роді еталон для тестування українських текстових подань, який охоплює 5 різних доменів.

Показана доцільність застосування процесу дистиляції задля пришвидшення моделей та встановлено обсяг даних, необхідний для дистиляції.

У п'ятому розділі проведена оцінка ефективності запропонованої технології обробки природномовних текстів у сфері управління персоналом. Зокрема, проведено оцінювання інформаційної технології інтенсифікації процесу відбору та ранжування резюме рекрутером. Проведено оцінювання технології аналізу резюме та зіставлення з вимогами вакансій в умовах відсутності рекрутера. Досліджено вплив окремих етапів на швидкість та якість інформаційної технології обробки природномовних текстів.

У *висновках* підсумовано виконані завдання дисертації, розкрито теоретичну та практичну цінність отриманих результатів, а також представлено інформацію щодо їх апробації та впровадження.

Наукова новизна отриманих результатів полягає у розробці та вдосконаленні методів обробки резюме та вакансій, зокрема:

– удосконалено модель подання токенів для візуального насичених документів, яка відрізняється від існуючих безпосередньою інтеграцією параметрів про стильові ознаки, що дозволяє підвищити якість подань токенів таких документів без використання методів комп'ютерного зору;

– вперше запропоновано метод подання фраз у контексті, що базується на використанні спеціальних маркерів для виділення фраз, які моделюються, що дозволяє значно пришвидшити процес подання фраз за рахунок використання лише однієї моделі та покращити якість в порівнянні з базовими методами;

– вперше запропоновано метод структурування документа для некерованого навчання подань текстів у сфері управління персоналом, що дозволяє адаптувати модель до домену та як наслідок підвищити якість сумаризації документів;

– удосконалено метод зменшення обсягу тексту, що оброблятиметься моделлю, який відрізняється від існуючих урахуванням структури документа та ключових фраз, для подальшої сумаризації, що дозволяє підвищити якість сумаризації довгих документів;

– вперше запропоновано модель векторизації документів у сфері управління людськими ресурсами на основі векторних подань секцій та механізму самоуваги разом із абсолютним позиційним кодуванням, що дозволяє покращити якість подань документів, довжина яких перевершує ліміт токенів, притаманний моделі.

Практичне значення отриманих результатів:

– адаптовано та вдосконалено технологію багатомовної дистиляції текстових подань Реймерса-Гуревич для текстів українською мовою, зокрема,

запропоновано використання функції втрат на основі косинусної подібності, а також встановлена сильна негативна кореляція між “90-м перцентилем розподілу косинусних коефіцієнтів подібності” моделі-вчителя та “середнім показником NMI”, досягнутим в процесі дистиляції векторних подань моделлю-студентом, що дозволяє надати рекомендації щодо тренування моделей при крос-лінгвістичній дистиляції знань. Створений еталонний набір даних для оцінювання якості векторних подань текстів українською мовою та викладений в публічний доступ на Гітхабі.

– Розроблено технологію для рекомендацій резюме в умовах відсутності рекрутера, яка була апробована в компанії Daxtra Technologies (Додаток Б – “Акт Впровадження технології використання штучних нейронних мереж для зіставлення вакансій та резюме”). Технологія базується на архітектурах типу “Трансформер” та поєднує етапи сегментації, парсингу та векторизації документів.

– Розроблено технологію використання штучних нейронних мереж для інтенсифікації процесу відбору та ранжування резюме рекрутером, яка була апробована в компанії Daxtra Technologies (Додаток Б – “Акт Впровадження технології використання штучних нейронних мереж для зіставлення вакансій та резюме”). В технологічну схему введено етап “сумаризація” для підвищення ефективності зіставлення вакансій та резюме.

– Розроблено алгоритм обробки природної мови, заснований на глибокому навчанні, для ефективної обробки візуально насичених документів, який був впроваджений в компанії Daxtra Technologies.

– Некерований метод навчання для векторних подань назв посад був впроваджений в компанії Daxtra Technologies, створений еталонний набір даних та викладений в публічний доступ на Гітхабі.

– Контекстно-освічений метод подання важливих фраз із застосуванням нової архітектури щодо представлення навичок кандидатів в домені управління персоналом був впроваджений в компанії Daxtra Technologies, та апробований в компанії Data Science UA (Додаток В – “Акт

апробації архітектури моделі векторного представлення фраз у контексті для їх групування та нормалізації”). Створений еталонний набір даних та викладений в публічний доступ на Гітхабі.

– Некерований метод навчання щодо тренування подань текстів у сфері управління персоналом був впроваджений в компанії Daxtra Technologies, створений еталонний набір даних та викладений в публічний доступ на Гітхабі.

Результати роботи впроваджені в науково-дослідній роботі “Методи, моделі, інформаційні технології розподілених систем підтримки прийняття організаційних рішень” (ДР 0121U111663). В навчальному процесі кафедри (дисципліна “Методи обробки текстів природної мови” Математичного забезпечення комп’ютерних систем використовується представлена у дисертаційній роботі архітектура контекстно-залежного подання фраз та алгоритм скорочення тексту з урахуванням структури документа та ключових фраз для сумаризації.

Ключові слова: обробка природної мови, штучні нейронні мережі, тонке налаштування, штучний інтелект, подання тексту, машинне навчання, нейронні мережі, глибоке навчання, мережі глибокого навчання, інформаційні технології, інтелектуальна система прогнозування, аналітичні методи, ефективність, математична модель, трансферне навчання.

ANNOTATION

Bocharova M.Y. Methods and information technology of subject-oriented analysis of natural language texts.

Dissertation for the degree of Doctor of Philosophy (PhD) in specialty 122 “Computer Science.” – I. I. Mechnikov Odesa National University, Odesa, 2025.

The dissertation presents the results of the research conducted by the applicant, which fulfills the urgent scientific task of creating models and methods for subject-oriented analysis of natural language texts, which is essential for the development of information technology.

The introduction of the dissertation substantiates the relevance of the research on the subject-oriented analysis of natural language texts, formulates the purpose, objectives and methods of the study, presents the scientific novelty and practical significance of the results obtained, and indicates the personal contribution of the applicant.

The first section of the paper investigates the current state of the problem of automated document analysis in the field of human resources management using artificial intelligence. It is shown that processing resumes to extract key information, matching vacancies and resumes is a necessary element in improving the efficiency of recruiting and a promising area for further improvement and development.

It is shown that the use of models that use computer vision is not appropriate for processing visually rich documents in the field of human resource management.

Based on the analysis of literature sources, the expediency of using context in modeling phrase representations is substantiated.

The problem of processing documents whose volume exceeds the limit of the models used for their processing is highlighted.

It is shown that the lack of data on the impact of the volume of training samples during their automatic generation (using large language models) on the quality of document summarization in the HR domain requires research in this direction. The approaches of unsupervised pre-training using the structure of

documents, as well as the loss functions used for pre-training (weighted loss function) need to be investigated.

The importance of the English language for the dissemination of the acquired knowledge on the use of artificial intelligence in recruiting is noted.

The need to improve the cross-linguistic distillation of vector representations to increase the efficiency of recruiting using artificial intelligence is determined.

The expediency of studying the impact of distillation on the speed and quality of the stages of natural language text processing technology for analyzing resumes and comparing them with job requirements has been established.

In the *second section*, we develop methods and models for subject-oriented processing of natural language texts. In particular, a new method of direct integration of parameters about style features is presented, where additional discrete features are vectorized and transferred to the Transformer architecture along with positional and token representations.

A new method of training job title representations based on the use of skill phrases specified in the job description is proposed. This method is based on the introduction of a special token to highlight and represent each skill, combined with contrast training to compare the average skill representation and job title from one job description.

A new method of unsupervised model training using document structure is proposed. In contrast to the traditional method, in which positive pairs for further contrast training are randomly selected from the document, the proposed method is based on the use of the document structure.

A new method of automatic creation of a job-resume dataset is proposed, which consists in using the document structure and a specific description of the last role and converting this record into a job description using a large language model.

A method of text reduction based on the document structure and key phrases is proposed. This method consists in shortening each section in proportion to its percentage contribution to the total length of the resume based on the selection of key phrases.

An unsupervised pre-training method was developed for summarizing documents in the field of human resources management. This method consists of using the “annotation” section of the resume for unsupervised training of the summarization model, as well as applying a weighted loss function that increases the weight for tokens that form key phrases.

The *3rd chapter* presents the information technology of subject-oriented analysis of natural language texts, which can be applied in two directions: making resume recommendations in the absence of a recruiter, and intensifying the process of selection and ranking of resumes by a recruiter, which allows recruiters to quickly and conveniently familiarize themselves with recommended candidates and filter them.

The presented technology is a sequence of the following stages: “Segmentation, Parsing, Summarization, Vectorization. As a result of applying these stages, the document is transformed into a set of attributes, annotations and vector representation, which are stored in a vector database.

It is shown that in order to evaluate the stages of the “AI ResJobFit” technology, it is necessary to calculate the following indicators: F1, Recall@N, MAP, MRR, nDCG, RougeN.

The fourth chapter substantiates the effectiveness and systematizes the developed methods for processing natural language texts in the field of human resource management.

It is established that the use of direct integration of parameters about style features (without the use of computer vision) allows to improve the quality of token classification in the task of segmenting resumes and vacancies, as well as extracting key information from them.

It is shown that a new learning method with context-aware phrase representation alignment leads to a significant improvement in the quality of phrase representations. It has been empirically determined that the “all negative pairs” approach to training in an asymmetric dataset using the multiple negative loss function for ranking is inappropriate, as it leads to a decrease in metrics.

It is shown that the use of a loss function based on cosine similarity leads to significant improvements (by 14.2% in terms of absolute NMI) compared to the use of the root mean square error loss function when distilling vector representations of texts from a well-trained teacher model.

A new and first-of-its-kind benchmark for testing Ukrainian text representations covering 5 different domains is presented.

The expediency of using the distillation process to speed up the models is shown and the amount of data required for distillation is determined.

The fifth chapter evaluates the effectiveness of the proposed technology for processing natural language texts in the field of human resources management. In particular, the information technology for intensifying the process of selecting and ranking resumes by a recruiter was evaluated. The technology for analyzing resumes and comparing them with the requirements of vacancies in the absence of a recruiter was evaluated. The influence of individual stages on the speed and quality of information technology for processing natural language texts is investigated.

The conclusions summarize the tasks of the dissertation, reveal the theoretical and practical value of the results obtained, and provide information on their testing and implementation.

The following *scientific results* were obtained as a result of the study:

– a model for representing tokens for visually rich documents has been improved, which differs from the existing ones by directly integrating parameters about style features, which allows improving the quality of token representations of such documents without using computer vision methods;

– for the first time, a method for representing phrases in context based on the use of special markers to highlight modeled phrases, which significantly speeds up the process of representing phrases by using only one model and improves the quality compared to basic methods;

– for the first time, a document structuring method for unsupervised learning of text representations in the field of human resources management is proposed,

which allows adapting the model to the domain and, as a result, improving the quality of document summarization;

– for the first time, a model of vectorization of documents in the field of human resources management based on vector representations of sections and a self-attention mechanism together with absolute positional coding is proposed, which allows to improve the quality of document submissions whose length exceeds the token limit inherent in the model;

– an improved method of reducing the volume of text, which differs from the existing ones by taking into account the structure of the document and key phrases, for further summarization, which allows to improve the quality of summarization of long documents;

Practical significance of the results.

– The Reimers-Gurevich technology of multilingual distillation of textual representations was adapted and improved for texts in Ukrainian, in particular, the use of a loss function based on cosine similarity was proposed, and a strong negative correlation was found between the “90th percentile of the distribution of cosine similarity coefficients” of the teacher model and the “average NMI” achieved in the process of vector representation distillation by the student model, which allows us to provide recommendations for model training in cross-linguistic knowledge distillation. A benchmark dataset for assessing the quality of vector representations of texts in Ukrainian was created and made publicly available on Github.

– A technology for recommending resumes in the absence of a recruiter was developed and tested at Daxtra Technologies (Appendix B - “Act of Implementation of the Technology of Using Artificial Neural Networks to Match Jobs and Resumes”). The technology is based on Transformer architectures and combines the stages of segmentation, parsing and vectorization of documents.

– A technology for using artificial neural networks to intensify the process of selecting and ranking resumes by a recruiter has been developed and tested at Daxtra Technologies (Appendix B - “Act of Implementation of the Technology for Using Artificial Neural Networks to Match Vacancies and Resumes”). The

technological scheme includes the “summarization” stage to increase the efficiency of job and resume matching.

– Developed a natural language processing model based on deep learning for efficient processing of visually rich documents, which was implemented at Daxtra Technologies.

– An unsupervised learning method for vector representations of job titles was implemented at Daxtra Technologies, a benchmark dataset was created and made publicly available on Github.

– A context-aware method of representing important phrases using a new architecture for representing candidate skills in the HR domain was implemented at Daxtra Technologies and tested at Data Science UA (Appendix B - “Act of testing the architecture of the model for vector representation of phrases in context for their grouping and normalization”). A benchmark dataset was created and made publicly available on Github.

– An unsupervised learning method for training text representations in the domain of human resources was implemented at Daxtra Technologies, a benchmark dataset was created and made publicly available on Github.

The results of the work are implemented in the research work “Methods, models, information technologies of distributed organizational decision support systems” (DR 0121U111663). In the educational process of the department (the discipline “Methods of Natural Language Processing” of the Mathematical Support of Computer Systems), the architecture of context-dependent phrase representation and the algorithm for text reduction, taking into account the structure of the document and key phrases for summarization, presented in the dissertation, are used.

Keywords: Natural Language Processing, Artificial Neural Networks, finetuning, Artificial Intelligence, text embeddings, machine learning, neural networks, deep learning, deep neural networks, Information Technology, intelligent forecasting system, Analytical methods, efficiency, mathematical model, transfer learning.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Наукові публікації у фахових виданнях України:

1. Bocharova M.Y., Malakhov E.V. and Mezhuyev V.I., VacancySBERT: the approach for representation of titles and skills for semantic similarity search in the recruitment domain. Applied Aspects of Information Technology. 6, 1 (Apr. 2023), 52–59. DOI: https://doi.org/10.15276/aait.06.2023.4 ISSN: 2617-4316

(Особистий внесок здобувача: запропоновано підхід до тренування подань назв посад, зібрано експериментальні дані та проведена оцінка результатів)

2. Bocharova M., Malakhov E., General-purpose text embeddings learning for Ukrainian language. Advanced Information Technology, vol.1(3), pp. 6–12, 2024 DOI: https://doi.org/10.17721/AIT.2024.1.01 ISSN: 2788-6603

(Особистий внесок здобувача: запропоновано удосконалення інформаційної теорії багатомовної дистиляції текстових подань на прикладі текстів українською мовою, орієнтуючись на визначення оптимальної функції втрат. Представлено новий датасет для тестування моделей векторних подань для української мови)

3. Bocharova M.Y., Malakhov E.V., ResJobFit - end-to-end artificial neural networks based technology for job-resume matching. Applied Aspects of Information Technology. 2024; Vol. 7, No. 4: 378–391. DOI: https://doi.org/10.15276/aait.07.2024.27 ISSN: 2617-4316

(Особистий внесок здобувача: запропоновано інформаційну технологію для зіставлення вакансій та резюме у векторному просторі)

Наукова публікація у виданнях, що входять до міжнародних наукометричних баз Scopus та Web of Science:

4. Bocharova, M., Malakhov, E. (2024). Improving information theory of context-aware phrase embeddings in HR domain. Eastern-European Journal of Enterprise Technologies, 5 (2 (131)), 53–60. https://doi.org/10.15587/1729-4061.2024.313970 (Scopus) Q3

(Особистий внесок здобувача: запропоновано удосконалення інформаційної теорії контекстно-освіченого подання фраз, яке базується на

використанні спеціальних токенів-маркерів для позначення меж фраз, створено датасет для оцінювання, проведено експерименти та оцінено результати)

Наукові праці, які засвідчують апробацію матеріалів дисертації:

5. Bocharova, M. Y., & Malakhov, E. V. (2024, February). CapStyleBERT: Incorporating capitalization and style information into BERT for enhanced resumes parsing. In Proceedings of the 2024 13th International Conference on Software and Computer Applications (pp. 249-254). DOI: https://doi.org/10.1145/3651781.3651820 (Scopus)

(Особистий внесок здобувача: запропоновано метод безпосереднього включення інформації про стильові ознаки в архітектуру типу "Трансформер", проведено огляд літератури та проаналізовано результати)

6. Bocharova, M., & Malakhov, E. (2022). Analysis of neural techniques for learning sentence representations. In Інформатика, інформаційні системи та технології: тези доповідей XIX Всеукр. конференції студентів і молодих науковців (pp. 114–116). Одеса, Україна.

(Особистий внесок здобувача: досліджено сучасний стан моделей щодо подань текстів у векторному просторі)

7. Bocharova, M. (2022). Trends and limitations in sentence embeddings learning. In 1st Student Scientific Conference of Joint Research Cooperation between Odesa I.I. Mechnikov National University and Huaiyin Institute of Technology (Online Edition), April 28–29, May 16, 2022.

(Особистий внесок здобувача: досліджено сучасний стан моделей щодо некерованого навчання подань текстів у векторному просторі)

8. Bocharova, M., & Malakhov, E. (2024). *Information technology in the AI-driven recruitment*. In *Тези доповідей, 76. Міжнародна науково-практична конференція "Сучасні інформаційні системи та технології в цифровому суспільстві"* (April 18–19, 2024). Харків, Україна.

(Особистий внесок здобувача: запропонована технологія виставлення вакансій та резюме та блоки, з яких ця технологія складається)

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ, УМОВНИХ ПОЗНАЧЕНЬ І ТЕРМІНІВ

NLP	Обробка природної мови (з англ. Natural Language Processing)
BP	Метод зворотного поширення помилки (з англ. Back Propagation)
RNN	Рекурентна нейронна мережа (з англ. Recurrent Neural Network)
BPE	Кодування байтових пар (з англ. Byte-Pair Encoding)
MSE	Середньоквадратична похибка (з англ. Mean Squared Error)
VDU	Візуальне розуміння документів (з англ. Visual Document Understanding)
NER	Розпізнавання іменованих сутностей (з англ. Named Entity Recognition)
BMM	Велика мовна модель (з англ. LLM — Large Language Model)
BIO	Схема тегування послідовностей (з англ. begin-inside-outside)

ВСТУП

В роботі розглянута проблема предметно-орієнтованого аналізу природномовних текстів. Відсутність ефективної технології обробки та зіставлення резюме та вакансій призводить до втрати рекрутерами часу на перегляд нерелевантних кандидатів. Це особливо актуально для україномовного сегменту.

Актуальність теми.

У сфері штучного інтелекту з акцентом на домені управління персоналом (HR) ефективний аналіз документів (вакансій та резюме) та семантичний пошук відіграє важливу роль у вдосконаленні процесів рекрутингу та управління кар'єрним розвитком. Динамічний характер ринків праці сучасного світу, зумовлений технологічними змінами, міграцією та цифровізацією, призводить до того, що велика кількість компаній стикається з необхідністю ефективного підбору, управління та розвитку персоналу.

Одним із ключових викликів для HR-спеціалістів є проблема зіставлення вимог вакансій з кваліфікацією кандидатів, що ускладнюється різнорідністю формату документів, використанням синонімів та неоднозначністю термінів. У цьому контексті штучний інтелект (ШІ) стає потужним інструментом для автоматизації аналізу та інтерпретації великих масивів даних, що містяться в резюме та описах вакансій. Інтелектуальні системи прогнозування дозволяють не лише швидше обробляти інформацію, але й забезпечують більш глибокий аналіз змісту завдяки можливостям семантичного пошуку та рекомендацій.

Семантичний пошук, як одна з передових технологій у сфері ШІ, надає можливість інтерпретувати зміст документів на глибшому рівні, враховуючи контекст і значення ключових термінів. Це відкриває нові можливості для підвищення точності підбору кандидатів на основі зіставлення не тільки ключових слів, а й змістовних відповідностей між вимогами вакансій та досвідом кандидата. Такий підхід значно перевищує можливості традиційних

систем пошуку за ключовими словами та пропонує більш релевантні результати, що сприяє оптимізації процесу найму.

Автоматизований аналіз навичок та компетенцій може допомогти організаціям визначити, якими навичками вже володіють їхні працівники, та виокремити ті, які можуть потребувати подальшого розвитку. Крім того, виявлення нових актуальних компетенцій, затребуваних на ринку праці, може сприяти більш цілеспрямованим ініціативам з підвищення кваліфікації, забезпечуючи конкуретноспроможність працівника на ринку праці.

Розробка інформаційної технології автоматизованого аналізу резюме та вакансій на основі вдосконалених та нових методів штучного інтелекту є актуальною, оскільки дозволяє значно підвищити ефективність рекрутингу.

Враховуючи швидкий темп змін на ринку праці та зростаючі вимоги до бізнесу, використання ШІ для аналізу великих масивів даних є важливим фактором для успішного розвитку компаній та працівників у майбутньому.

Дана робота зосереджена на вивченні теоретичних питань та створенню нових методів та інформаційної технології для обробки природної мови з застосуванням моделей архітектури типу “Трансформер”.

Вирішення цих питань дозволить забезпечити більше точну обробку природномовних текстів, сформувати теоретичну базу для застосування конкурентоспроможних систем, в чому зацікавлена Україна та міжнародне співтовариство, і тому є особливо актуальним.

Результати роботи були впроваджені в науково-дослідній роботі “Методи, моделі, інформаційні технології розподілених систем підтримки прийняття організаційних рішень” (ДР 0121U111663).

Метою дисертаційної роботи є підвищення ефективності рекрутингу за рахунок інтенсифікації процесів відбору, ранжування резюме та вироблення рекомендацій шляхом створення методів та інформаційної технології предметно-орієнтованого аналізу природномовних текстів.

В якості критерія ефективності технології розглядається швидкість аналізу резюме та вакансій та релевантність рекомендацій резюме щодо вимог вакансій при мінімізації впливу людського фактору.

Для досягнення поставленої мети поставлено наступні задачі:

1. Проаналізувати сучасні методи штучного інтелекту та машинного навчання в обробці природномовних документів та визначити шляхи оптимізації процесу прийняття рішень у підборі персоналу на основі аналізу великих обсягів даних.

2. Запропонувати модель представлення візуально насичених документів без використання методів комп'ютерного зору.

3. Запропонувати підхід до тренування подань назв посад, що базується на використання фраз-навичок, які містяться в описі роботи, та створити еталонний набір даних.

4. Розробити технологію виділення фраз для їх групування та нормалізації та створити еталонний набір даних.

5. Розробити метод некерованого навчання моделі подання текстів у векторному просторі на основі врахування структури документів.

6. Розробити метод векторизації структурованих документів в домені управління персоналом

7. Розробити метод зменшення обсягу тексту з урахуванням структури документа та ключових фраз.

8. Розробити метод некерованого попереднього тренування та застосування зваженої функції втрат для сумаризації документів з використанням їх структури. Визначити поріг кількості тренувальних зразків, який забезпечує підвищення продуктивності.

9. Удосконалити інформаційну теорію багатомовної дистиляції текстових подань на прикладі текстів українською мовою, орієнтуючись на визначення оптимальної функції втрат.

10. Розробити інформаційну технологію аналізу резюме та зіставлення з вимогами вакансій в умовах відсутності рекрутера на основі сегментації, парсингу і векторизації документів.

11. Розробити інформаційну технологію інтенсифікації процесу відбору та ранжування резюме рекрутером.

12. Визначити шляхи вдосконалення інформаційної технології предметно-орієнтованого аналізу текстів для підвищення якості зіставлення вакансій і резюме та скорочення часу, необхідного рекрутеру на ознайомлення з резюме.

Об'єктом дослідження є процес обробки природної мови в документах, пов'язаних з рекрутингом та управлінням персоналом, на кшталт резюме та вакансій.

Предметом дослідження є методи штучного інтелекту на основі глибокого навчання, методи ідентифікації та екстракції даних у візуально насичених документах, моделі представлення візуально насичених документів, контекстно-освічені методи подання фраз, методи звичайної та крос-лінгвістичної дистиляції векторних подань.

Методи дослідження

До методів, використаних в роботі, належать наступні: методи машинного навчання, зокрема, глибоке навчання, теорія ймовірності та статистики; метод контекстно-освіченого подання фраз та текстів у векторному просторі; методи подання токенів візуально насичених документів у векторному просторі з урахуванням їх візуальних характеристик; методи некерованого навчання для вивчення подань текстів у векторному просторі; методи звичайної та крос-лінгвістичної дистиляції векторних подань токенів та текстів; інструменти NLP.

Експериментальні дослідження проведені з використанням класичних та оригінальних методик, проаналізовано відповідність отриманих результатів даним інших авторів. До отриманих даних застосовували системний підхід, аналіз, синтез.

Розроблені нові методи та технології не суперечать існуючим теоріям у сфері машинного навчання та обробки природної мови.

Наукова новизна отриманих результатів полягає у розробці та вдосконаленні методів обробки резюме та вакансій, зокрема:

– удосконалено модель подання токенів для візуального насичених документів, яка відрізняється від існуючих безпосередньою інтеграцією параметрів про стильові ознаки, що дозволяє підвищити якість подань токенів таких документів без використання методів комп'ютерного зору;

– вперше запропоновано метод подання фраз у контексті, що базується на використанні спеціальних маркерів для виділення фраз, які моделюються, що дозволяє значно пришвидшити процес подання фраз за рахунок використання лише однієї моделі та покращити якість в порівнянні з базовими методами;

– вперше запропоновано метод структурування документа для некерованого навчання подань текстів у сфері управління персоналом, що дозволяє адаптувати модель до домену та як наслідок підвищити якість сумаризації документів;

– удосконалено метод зменшення обсягу тексту, що оброблятиметься моделлю, який відрізняється від існуючих урахуванням структури документа та ключових фраз, для подальшої сумаризації, що дозволяє підвищити якість сумаризації довгих документів;

– вперше запропоновано модель векторизації документів у сфері управління людськими ресурсами на основі векторних подань секцій та механізму самоуваги разом із абсолютним позиційним кодуванням, що дозволяє покращити якість подань документів, довжина яких перевершує ліміт токенів, притаманний моделі.

Практичне значення отриманих результатів:

– адаптовано та вдосконалено технологію багатомовної дистиляції текстових подань Реймерса-Гуревич для текстів українською мовою, зокрема, запропоновано використання функції втрат на основі косинусної подібності, а

також встановлена сильна негативна кореляція між “90-м перцентилем розподілу косинусних коефіцієнтів подібності” моделі-вчителя та “середнім показником NMI”, досягнутим в процесі дистиляції векторних подань моделлю-студентом, що дозволяє надати рекомендації щодо тренування моделей при крос-лінгвістичній дистиляції знань. Створений еталонний набір даних для оцінювання якості векторних подань текстів українською мовою та викладений в публічний доступ на Гітхабі.

– Розроблено технологію для рекомендацій резюме в умовах відсутності рекрутера, яка була апробована в компанії Daxtra Technologies (Додаток Б – “Акт Впровадження технології використання штучних нейронних мереж для зіставлення вакансій та резюме”). Технологія базується на архітектурах типу “Трансформер” та поєднує етапи сегментації, парсингу та векторизації документів.

– Розроблено технологію використання штучних нейронних мереж для інтенсифікації процесу відбору та ранжування резюме рекрутером, яка була апробована в компанії Daxtra Technologies (Додаток Б – “Акт Впровадження технології використання штучних нейронних мереж для зіставлення вакансій та резюме”). В технологічну схему введено етап “сумаризація” для підвищення ефективності зіставлення вакансій та резюме.

– Розроблено алгоритм обробки природної мови, заснований на глибокому навчанні, для ефективної обробки візуально насичених документів, який був впроваджений в компанії Daxtra Technologies.

– Некерований метод навчання для векторних подань назв посад був впроваджений в компанії Daxtra Technologies, створений еталонний набір даних та викладений в публічний доступ на Гітхабі.

– Контекстно-освічений метод подання важливих фраз із застосуванням нової архітектури щодо представлення навичок кандидатів в домені управління персоналом був впроваджений в компанії Daxtra Technologies, та апробований в компанії Data Science UA (Додаток В – “Акт апробації архітектури моделі векторного представлення фраз у контексті для

їх групування та нормалізації”). Створений еталонний набір даних та викладений в публічний доступ на Гітхабі.

– Некерований метод навчання щодо тренування подань текстів у сфері управління персоналом був впроваджений в компанії Daxtra Technologies, створений еталонний набір даних та викладений в публічний доступ на Гітхабі.

– Створений еталонний набір даних для оцінювання векторних подань текстів українською мовою та викладений в публічний доступ на Гітхабі.

Результати роботи впроваджені в науково-дослідній роботі “Методи, моделі, інформаційні технології розподілених систем підтримки прийняття організаційних рішень” (ДР 0121U111663). В навчальному процесі кафедри (дисципліна “Методи обробки текстів природної мови” Математичного забезпечення комп’ютерних систем використовується представлена у дисертаційній роботі архітектура контекстно-залежного подання фраз та алгоритм скорочення тексту з урахуванням структури документа та ключових фраз для сумаризації.

Особистий вклад здобувача

Наукові положення і результати, що представлені в дисертаційній роботі, отримані здобувачем особисто. У наукових роботах, написаних у співавторстві, здобувачеві належить:

1) фахові видання України:

[1] – здобувачем запропоновано метод безпосереднього включення інформації про стильові ознаки в архітектуру типу “Трансформер”, проведено огляд літератури та проаналізовано результати; співавтором Є.В. Малаховим – концептуалізація та формулювання завдання;

[2] – здобувачем запропоновано підхід до тренування подань назв посад, зібрано експериментальні дані та проведена оцінка результатів; співавтором І. Є.В. Малаховим – концептуалізація, формулювання завдань, ідея моделі; співавтором В.І. Межуєвим – огляд літератури.

[3] – здобувачем запропоновано удосконалення інформаційної теорії контекстно-освіченого подання фраз, яке базується на використанні спеціальних токенів-маркерів для позначення меж фраз, створено датасет для оцінювання, проведено експерименти та оцінено результати; співавтором Є.В. Малаховим – концептуалізація, формулювання завдання, методологія, ідея моделі;

[4] – здобувачем запропоновано удосконалення інформаційної теорії багатомовної дистиляції текстових подань на прикладі текстів українською мовою, орієнтуючись на визначення оптимальної функції втрат. Представлено новий датасет для тестування моделей векторних подань для української мови; співавтором Є.В. Малаховим – концептуалізація, формулювання завдання, методологія, ідея визначення оптимальної функції втрат;

2) матеріали апробаційного характеру:

[5] – здобувачем запропоновано інформаційну технологію для зіставлення вакансій та резюме у векторному просторі; співавтором Є.В. Малаховим – концептуалізація, формулювання завдання, методологія, ідея інформаційної технології;

[6] – здобувачем досліджено сучасний стан моделей щодо подань текстів у векторному просторі; співавтором Є.В. Малаховим – концептуалізація, формулювання завдання;

[7] – здобувачем досліджено сучасний стан моделей щодо некерованого навчання подань текстів у векторному просторі; співавтором Є.В. Малаховим – концептуалізація, формулювання завдання;

[8] – здобувачем запропонована технологія виставлення вакансій та резюме та блоки, з яких ця технологія складається; співавтором Є.В. Малаховим – концептуалізація, формулювання завдання, методологія;

Апробація результатів дисертації.

Матеріали досліджень були заслухані на конференціях:

“1st Student Scientific Conference of Joint Research Cooperation between Odesa I.I. Mechnikov National University and Huaiyin Institute of Technology” (Online, 2022);

“2024 13th International Conference on Software and Computer Applications (ICSCA 2024)” (Bali Island, Indonesia, February 1-3, 2024). Індексована в Scopus.

Міжнародна науково-практична конференція “Сучасні інформаційні системи та технології в цифровому суспільстві”. Харків, 18 - 19 квітня 2024 р.

Публікації

За результатами досліджень опубліковано 8 робіт, в тому числі 3 статті українських періодичних фахових виданнях, 1 стаття в журналі, проіндексованому в науковій базі Скопус.

Структура та обсяг дисертації. Дисертаційна робота складається з анотацій, змісту, вступу, п’яти розділів, висновків, списку використаних джерел та додатків. Загальний обсяг дисертації складає 188 сторінок, з них основного тексту – 148; 31 рисунок по тексту; 21 таблиця по тексту; список використаних джерел на 11-ти сторінках; додатки на 8-и сторінках.

1 АНАЛІЗ ЗАСТОСУВАНЬ ОБРОБКИ ПРИРОДНОЇ МОВИ ДЛЯ ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ РЕКРУТИНГУ

1.1 Аналіз ринку та загальна характеристика рекрутингу на основі штучного інтелекту

1.1.1 Загальна характеристика рекрутингу на основі штучного інтелекту

Сфера управління людськими ресурсами (HR) є ключовою складовою стратегічного розвитку компаній, оскільки ефективне управління персоналом безпосередньо впливає на конкурентоспроможність бізнесу. Одним з важливих напрямків діяльності у сфері управління людськими ресурсами є рекрутинг. Під рекрутингом розуміють комплексний процес пошуку, залучення, відбору та найму кваліфікованих кандидатів на вакантні посади [9]. Рекрутинг на основі штучного інтелекту відносять до застосування технологій і алгоритмів штучного інтелекту в процесах відбору і найму персоналу.

Рекрутинговий процес поділяють на наступні етапи [9]:

1) Створення вичерпного профілю кандидата.

2) Пошук кандидата (компанія розміщує вакансію, а кандидати звертаються до компанії / сама компанія вибирає та контактує кандидатів, використовуючи загальнодоступні професійні профілі).

3) Відбір — перегляд HR-фахівцями резюме та супровідних листів, проведення онлайн-інтерв’ю, співбесід, оцінки тестового завдання, перевірки рекомендацій тощо.

4) Найм на роботу

Оцінювання навичок кандидатів на етапі відбору виділено в окрему операцію, що пояснюють її важливістю [10].

Традиційно пошук і залучення нових талантів завжди був “людською” професією, наповненою безліччю повторюваних завдань. Сьогодні штучний

інтелект (ШІ) та інші інструменти рекрутингу, такі як чат-боти, можуть виконувати близько 80% цієї адміністративної роботи [11].

Інтеграція технологій штучного інтелекту і машинного навчання розширює можливості систем для онлайн-рекрутингу, надаючи такі функції, як інтелектуальна система прогнозування щодо підбору кандидатів, прогнозна аналітика і автоматизовані відповіді. Алгоритми ШІ можуть швидко обробляти величезні масиви даних, надаючи змогу компаніям швидше і точніше виявляти кандидатів, які найкраще відповідають вимогам. Це не тільки оптимізує процес пошуку працівників, а й підвищує якість найму.

Інструменти штучного інтелекту націлені на наступні стадії рекрутингу [10]:

1. Підбір персоналу: швидкий пошук і зв'язок з талановитими фахівцями
2. Відбір: швидке визначення найкращих кандидатів
3. Проведення співбесід: полегшення дистанційного найму та економія часу

Інструменти відбору ШІ варіюються від аналізу резюме до оцінки поведінки та навичок [11].

Етап відбору кандидатів є найбільш тривалою стадією [9], що зумовлено необхідністю проаналізувати значну кількість резюме та зіставити з вимогами до кандидатів. Тому розробка технологій на основі ШІ для підвищення швидкості та якості саме цього етапу сприятиме підвищенню ефективності рекрутингу в цілому.

Інструменти для автоматичної перевірки резюме зазвичай належать до однієї з трьох категорій [9]:

1. На основі ключових слів: штучний інтелект шукає ключові слова, фрази та шаблони в тексті для сортування кандидатів.

2. Граматичні: алгоритми машинного навчання використовують список попередньо визначених граматичних правил, розбиваючи слова і фрази в резюме, для правильної обробки значення кожного речення.

3. Статистичні: моделі на основі штучних нейронних мереж в процесі тренування навчають моделювати семантичні залежності між словами та фразами в резюме. Такі моделі можуть враховувати не лише окремі ключові слова, а й їхнє значення у різних контекстах, що підвищує точність відбору кандидатів.

Автоматизація в рекрутингу досягла середнього рівня розвитку, надаючи рекрутерам інструменти для спрощення та оптимізації процесу найму. Постійний розвиток технологій і нові можливості можуть призвести до їх значного поліпшення. Аналіз резюме може бути як повністю так і частково автоматизованим шляхом застосування ШІ [9]. Повністю автоматизована операція аналізу резюме та зіставлення вимог до кандидата з даними резюме виключає присутність рекрутера на цій операції етапу відбору кандидатів, що дає змогу знизити суб'єктивність у процесі найму та знизити витрати. Однак, сучасні розробки в цій галузі не дають достатнього ефекту в співвідношенні швидкості та якості результатів [10], що показує необхідність у подальшому удосконаленні та розробленні нових технологій ШІ в цьому напрямку.

Окремим важливим завданням для рекрутинга є пошук талантів. На сучасному ринку праці, що швидко змінюється, ефективне та результативне залучення талантів є ключовим викликом для компаній. Інтенсивна конкуренція за найкращі таланти в поєднанні з великою кількістю заявок на відкриті вакансії призвела до нових складнощів в обробці та відборі профілів кандидатів [10]. Щоб впоратися зі швидкозмінним бізнес-середовищем і зберегти конкурентні переваги, для компаній критично важливо переосмислити, як приймати рішення, пов’язані з талантами, у кількісному форматі. Конкуренція за таланти змушує компанії використовувати інноваційні методи рекрутингу, зокрема, персоналізовані рекомендації

кандидатів та аналіз профілів у соціальних мережах з використанням технологій ШІ.

Доступність даних в сфері управління персоналом відкриває нові можливості для бізнес-лідерів. У цьому контексті, як новий напрям прикладної науки про дані в управлінні людськими ресурсами, аналіз талантів привернув значну увагу як з боку академічних кіл, так і з боку індустрії. Зокрема, аналіз талантів, також відомий як аналіз робочої сили або аналітика людей (“workforce analysis” або “people analytics”), зосереджується на використанні технологій машинного та, зокрема, глибокого навчання. На практиці, аналіз робочої сили відіграє вирішальну роль в стратегічному управлінні людськими ресурсами, включаючи рекрутинг.

Таким чином, автоматизований аналіз документів в сфері управління персоналом, в тому числі обробка резюме для витягнення ключової інформації, а також зіставлення вакансій та резюме є необхідним елементом підвищення ефективності рекрутингу і перспективним напрямком для подальшого його удосконалення і розвитку.

1.1.2 Аналіз ринку рекрутингу на основі штучного інтелекту

Оцифрування процесів управління персоналом стало важливою тенденцією останнього десятиліття. Компанії активно впроваджують сучасні HRM-системи (з англ. Human Resource Management Systems), такі як SAP SuccessFactors, Workday, BambooHR, які автоматизують рекрутингові процеси, оцінку продуктивності та управління кар'єрним розвитком працівників [11].

За підсумками 2023 року обсяг світового ринку програмного забезпечення для онлайн-рекрутингу досяг 661,56 мільйонів доларів. Роком раніше витрати в цьому секторі оцінювали в 612,5 мільйонів доларів [12]. Аналітики Market Research Future вважають, що надалі показник середньорічного темпу зростання в складних відсотках становитиме 6,8%.

Очікується, що в Азійсько-Тихоокеанському регіоні спостерігатиметься пришвидшене зростання на рівні 7,01 %. Таким чином, до 2032 року глобальні витрати в цьому секторі можуть досягти 1,2 млрд доларів [12].

Автори дослідження поділяють ринок програмного забезпечення для онлайн-рекрутингу на засоби відстеження кандидатів, маркетингові інструменти, онбордінг (поетапне залучення фахівця в соціальне і робоче середовище компанії), засоби залучення кандидатів через рекомендації та інструменти оцінки. У 2023 році витрати в цих сегментах становили 197,7; 123,6; 98,9; 142,9 та 49,4 мільйонів доларів відповідно [12].

До списку провідних гравців галузі входять: SAP SuccessFactors, Glassdoor, Workday, Indeed, iCIMS, Bullhorn, SmartRecruiters, LinkedIn, Jobvite, ADP, Monster, CareerBuilder, Oracle Taleo, ZipRecruiter.

У 2023 році Північна Америка лідирувала з оцінкою витрат у розмірі 247,2 мільйонів доларів, що відображає значний попит у секторі рекрутингу, зумовлений впровадженням передових технологій і зрілим ринком праці. Далі йде Європа з витратами на рівні 173,0 мільйонів доларів. В Азіатсько-Тихоокеанському регіоні витрати на рекрутингові системи становили близько 133,5 мільйонів доларів. На Південну Америку припало 44,5 мільйонів доларів, на Близький Схід і Африку — 14,3 мільйонів доларів (рисунок. 1.1) [12].

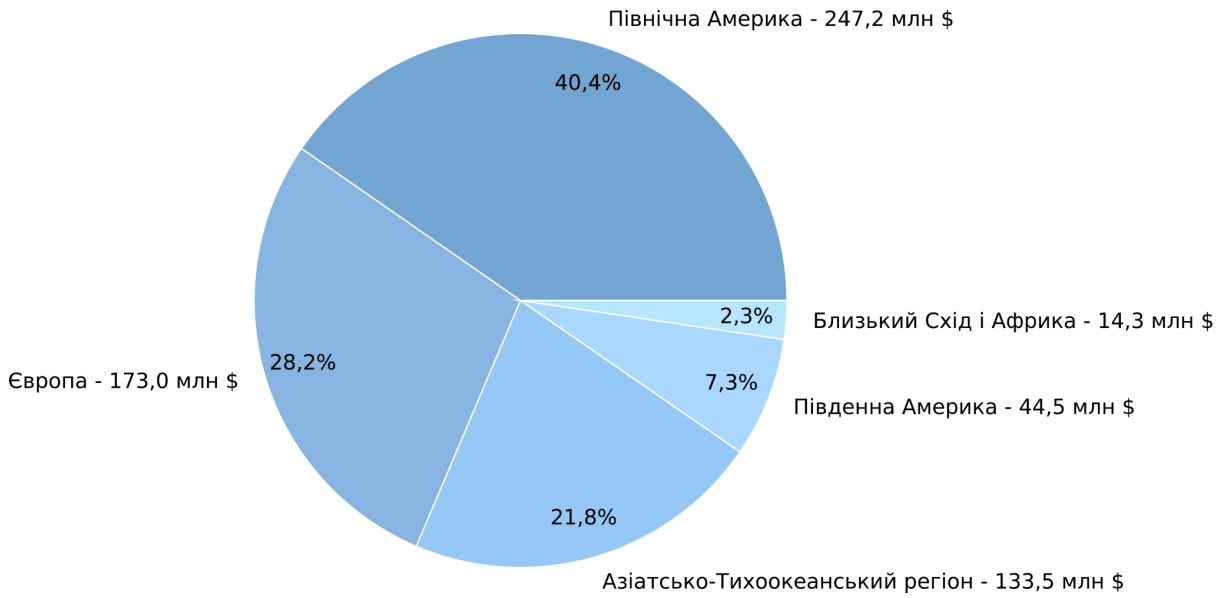


Рисунок 1.1 — Розподіл витрат в секторі рекрутингу у світі [12]

Кількість рекрутингових агентств складає більше 160 000, причому більше 25 000 та 35 000 знаходяться у США та Великобританії відповідно. Тобто на ці дві країни припадає 60 000 з 160 000 рекрутингових ангенств, що складає 37,5% від світової кількості [12]. З цього випливає важливість саме англійської мови для поширення набутих знань щодо використання штучного інтелекту у рекрутингу.

1.2 Еволюція мовних моделей

Обробка природної мови (з англ. Natural Language Processing, NLP) — це міждисциплінарна галузь інформатики та штучного інтелекту, яка займається взаємодією між комп'ютерами та людськими мовами. Основна мета NLP полягає в тому, щоб навчити машини розуміти, інтерпретувати та генерувати природну мову. Це включає широкий спектр завдань, таких як інформаційний пошук, аналіз тексту, машинний переклад, розпізнавання іменованих сутностей, сумаризація, відповіді на запитання тощо [13, 14]. Обробка природної мови є ключовим елементом сучасних інтелектуальних

систем прогнозування і використовується у пошукових системах, чат-ботах, голосових помічниках і багатьох інших технологіях [15, 16].

Мовні моделі є центральним компонентом обробки природної мови, які дозволяють комп'ютерам розуміти та генерувати текстові дані. Вони будуються на основі статистичних та нейронних підходів, що аналізують мовні структури, створюють прогнози щодо наступного слова або фрази, а також розпізнають складні мовні закономірності. Розвиток мовних моделей значно вплинув на вдосконалення таких завдань, як автоматичний переклад, сумаризація текстів, діалогові чи пошукові системи тощо [17, 18].

1.2.1 Традиційні мовні моделі

Традиційні мовні моделі, які застосовували для лінгвістичного аналізу тексту, розглядали мову як текстову послідовність, і задачею було передбачення наступного слова чи фрази, спираючись на послідовності попередніх слів. Це виступило мотивацією так званих N-gram моделей.

Так, маючи послідовність слів $\{w_1, w_2, \ldots, w_N\}$, традиційна мовна модель намагається максимізувати ймовірність $P(w_i) = P(w_1, w_2, \ldots, w_{i-1})$ спираючись на ланцюг умовних ймовірностей, що можна виразити математичною моделлю:

$$P(w) = \prod_{i=1}^{N} P(w_i | w_{i-1}, w_{i-2}, \ldots w_1), \quad (1.1)$$

де w_i — слово на *i*-ій позиції в текстовій послідовності, *P* - ймовірність, *N* зазвичай встановлювалося в межах 2-5.

Тоді ймовірність слова $P(w_1)$, спираючись на послідовності в найпростішому вигляді можна виразити наступною формулою:

$$P(w_i | w_{i-2}, w_{i-2}, \ldots w_1) \approx \frac{Count(w_{i-2}, w_{i-2}, w_1)}{Count(w_{i-2}, w_{i-2})}, \quad (1.2)$$

де $Count$ — кількість послідовностей такого типу в оригінальному корпусі.

Головною проблемою такого підходу виступає розрідженість даних, деякі послідовності можуть взагалі жодного разу не зв’явитися в тренувальному наборі даних. Для подолання цієї проблеми були запроваджені методи згладжування [19].

Але навіть при використанні згладжування Катца N-gram моделі стикаються з проблемою надзвичайно розріджених даних. Так, якщо взяти послідовність довжиною N = 4 та словник розміру 32,000 (що є дуже малим за сучасними стандартами), система буде змушена зберігати $30{,}000^{(4-1)} = 27 * 10^{12}$ комбінацій з відповідними вірогідностями, що не є реальним. Задля подолання проблеми була розроблена N-gram модель, заснована на класах [20]. В цьому підході, словник розміру V спочатку поділяється на K класів з використанням спеціальної функції. Ця функція забезпечує розподілення слів без перекриття, а кількість класів вибирається емпірично.

В дослідженні [20] показуються, що слова, схожі за сенсом, класифіковані в однакові класи. Це перша мовна модель, яка здатна виражати семантичну структуру, перевершуючи класичну мету передбачення наступного слова спираючись на послідовність попередніх слів.

Тому N-gram моделі та побудовані на класах N-gram моделі розглядаються як перші вдалі традиційні мовні моделі. Але досягнення традиційних моделей обмежені, тому що статистичний аналіз та правила не здатні опрацювати складні аспекти людської мови. Першим стрибком в обробці природномовних текстів стало запровадження моделей, заснованих на глибокому навчанні, які будуть розглянуті нижче.

1.2.2 Мовні моделі, засновані на глибокому навчанні

В цій секції будуть розглянуті базова структура нейронної мережі, включаючи рекурентні нейронні мережі, мережі LSTM та концепція самоуваги і моделі класу “Трансформер”.

Штучна нейронна мережа представляє собою обчислювальні блоки, схожі на біологічні нейрони, як показано на рисунку 1.2.

Кожна w_i — це вага дендрита, що вказує на силу зв’язку цього вузла з попереднім вузлом через дендрит i. Тоді, враховуючи вхідні дані $\{x_i\}$ від попередніх вузлів, тіло “клітини” виконує лінійне обчислення за формулою $out=\sum_i w_i x_i+b$, де b — це зміщення (bias) цієї клітини. Нарешті, ця сума подається до нелінійної “функції активації” f, яка визначає вихід вузла.

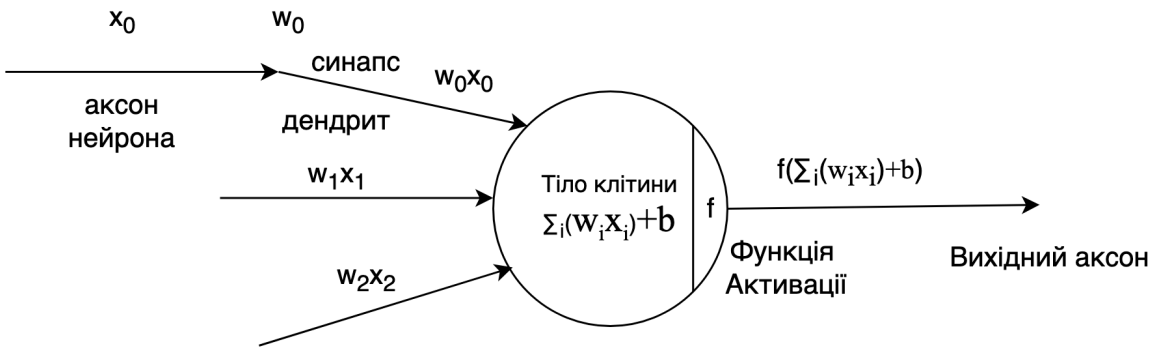


Рисунок 1.2 — Схема “клітини” штучної нейронної мережі [19]

Сигмоїдна функція $S(x)=1 / (1 + e^{-x})$ та гіперболічна тангенс-функція $tanh(x)= (e^{x} - e^{-x}) / (e^{x} + e^{-x})$ є дуже популярними функціями активації.

Правильний вибір функції активації f є критично важливим для ефективності нейронної мережі, оскільки нелінійна (високого порядку) здатність обробки всієї мережі значною мірою залежить від властивостей функції f. Тобто, враховуючи вхідні дані x, вхідний шар передає їх до прихованого шару (або шарів). Потім вузли прихованого шару (шарів) виконують лінійне обчислення та нелінійну активацію, як обговорено вище, перш ніж передати результат до вихідного шару. Після цього результат y

порівнюється з цільовою міткою $y^\wedge$ за допомогою функції помилки E. Помилка — це різниця між цільовим значенням і результатом.

За допомогою методу зворотного поширення помилки (Back Propagation, BP) помилка передається назад через мережу, шар за шаром, щоб скоригувати ваги всіх прихованих вузлів. Це зворотне поширення є реалізацією правила ланцюга для часткових похідних. Таким чином, обчислювальна складність $\partial E/\partial w_{ij}$ зменшується з експоненційної до лінійної. Типова формула функції помилки E та формула градієнтного спуску для оновлення кожної ваги під час зворотного поширення виглядають так:

$$E = \frac{1}{2}(y_i^\wedge - y_i)^2, \tag{1.3}$$

$$w_{ij} \leftarrow w_{ij} - \eta\frac{\partial E}{\partial w_{ji}}, \tag{1.4}$$

де η — це швидкість навчання мережі (learning rate). Інтуїтивно градієнтний спуск базується на тому, наскільки кожна вага впливає на загальну помилку, і кожна вага коригується відповідно. Після багатьох ітерацій помилка (різниця між y і $y^\wedge$) буде значно зменшена. Таким чином, мета нейронної мережі полягає в тому, щоб точно і правильно відобразити вхідні дані x на цільове значення $y^\wedge$.

Однак для задач моделювання послідовностей типова структура нейронної мережі прямого поширення не є достатньою. Це пов’язано з тим, що у вхідних даних міститься часовий контекст, який неможливо проаналізувати мережею без внутрішніх зв’язків між шарами. Для того щоб врахувати часову інформацію, до мережі додаються внутрішньошарові зв’язки, що перетворює її на рекурентну нейронну мережу (Recurrent Neural Network, RNN) [19]. На рисунку 1.5 показано базову структуру RNN.

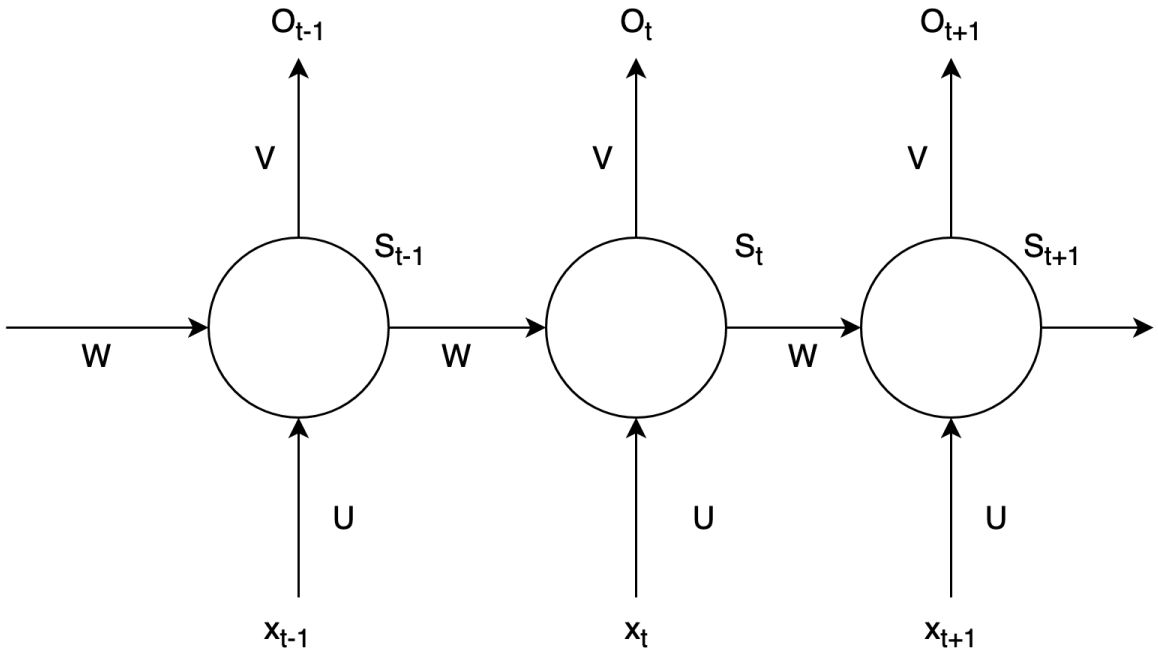


Рисунок 1.5 — Рекурентна нейронна мережа [14]

Тут $\{x_i\}$ та $\{o_i\}$ є відповідно вхідною та вихідною послідовностями, тоді як $\{s_i\}$ представляє послідовність прихованого шару. Однак, на практиці вхідними даними не буде вся навчальна послідовність $\{x_1,\ldots,x_i\}$. Замість цього навчання RNN відбуватиметься в багатьох кроках. На кожному кроці формується батч даних довжини N, $\{x_1,\ldots,x_N\}$, який є послідовним сегментом із усієї послідовності.

Матриці U, V і W використовуються для представлення ваг, а f і g задаються як функції активації. Тоді формули прямого поширення виглядають так:

$$o_i = g(Vs_i), \tag{1.5}$$

$$s_i = f(Ux_i + Ws_{i-1}), \tag{1.6}$$

для всіх i = 1, …, N. Якщо застосувати цю формулу рекурсивно, то отримаємо:

$$o_i = g(Vs_i)$$

$$= g(Vf(Ux_i + Ws_{i-1})) = g(Vf(Ux_i + Wf(Ux_{i-1} + Ws_{i-2}))) , \tag{1.7}$$

Таким чином, інформація з попередніх часових кроків передається до поточного кроку і впливає на вихід o_i. Крім того, ця архітектура забезпечує легке додавання додаткових шарів (звідси й походить концепція “глибини”). У цьому випадку вихідна послідовність $\{o_i\}_{i=1..N}$ поточного шару буде виступати в ролі вхідної послідовності $\{x_i\}_{i=1..N}$ для наступного шару.

Однак існує серйозна проблема класичної рекурентної нейронної мережі: помилка під час зворотного поширення може експоненціально зростати (або зменшуватися). Це явище називається вибухом градієнта (gradient exploding) або згасанням градієнта (gradient vanishing) [19]. Для вирішення цієї проблеми Хохрейтер і Шмідхубер у 1997 році [19] створили мережу довготривалої короткочасної пам'яті (Long Short-Term Memory, LSTM), яка представляє собою модифікацію RNN. Незабаром після цього було удосконалено архітектуру LSTM шляхом додавання механізму вибіркового збереження інформації з попередніх часових кроків у кожній LSTM-комірці [19]. Це вдосконалення з того часу набуло поширення. Схематично LSTM зображена на рисунку 1.5.

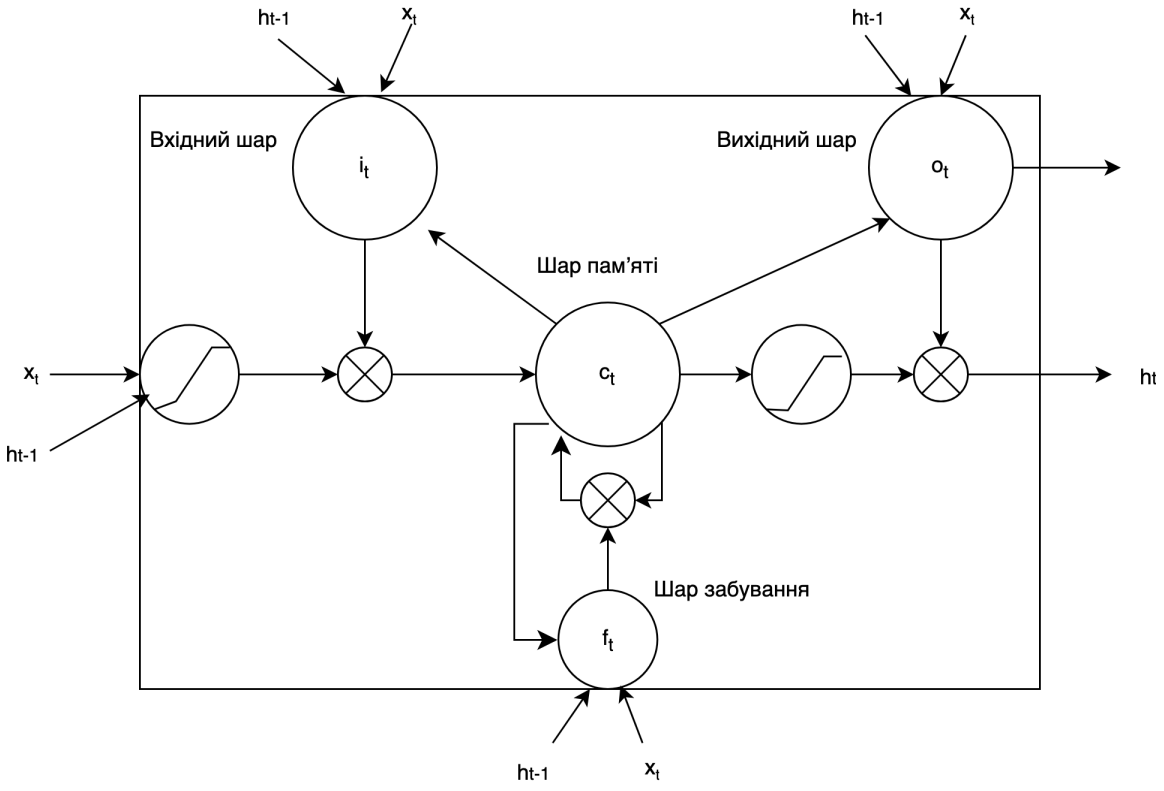


Рисунок 1.5 — Структура вдосконаленої LSTM-комірки [19]

Математичне визначення операції LSTM-комірки, яке описане в [19]:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i), \quad (1.7)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f), \quad (1.8)$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \quad (1.9)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o), \quad (1.10)$$

$$h_t = o_t tanh(c_t), \quad (1.11)$$

Тут σ представляє сигмоїдну функцію: $\sigma(x)=1 / (1 + e^{-x})$. Як показано на рисунку 1.5, у комірці LSTM є три шлюзи: вхідний шлюз i_t, шлюз забування f_t і вихідний шлюз o_t. Ці шлюзи працюють разом, щоб зберігати лише цінну інформацію в кожній комірці, як описано далі. Навчальні матриці ваг позначаються W із відповідним індексом. Наприклад, матриця ваг для вхідного сигналу x_t, коли він потрапляє до шлюзу забування f_t, позначається як W_{xf}, а матриця ваг для прихованого стану h_{t-1}, коли він використовується для обчислення пам'яті c_t позначається як W_{hc}.

Оновлення довготривалої пам’яті є рекурентним процесом: на кожному часовому кроці t, базуючись на вхідному сигналі x_t та прихованому стані h_{t-1}, LSTM-комірка спочатку обчислює значення свого вхідного шлюзу i_t і шлюзу забування f_t відповідно до рівнянь (1.7) та (1.8). Потім пам’ять c_{t-1} оновлюється до c_t відповідно до рівняння (1.9). Тобто, значення шлюзу забування f_t визначає, скільки довготривалої пам’яті з попередніх кроків зберігається, тоді як значення вхідного шлюзу i_t визначає, скільки інформації з поточного вхідного сигналу x_t та прихованого стану h_{t-1} потрапляє до пам’яті. Нарешті, вихід o_t та новий прихований стан h_t обчислюються відповідно до рівнянь (1.10) та (1.11), базуючись на x_t, h_{t-1} та новій довготривалій пам’яті c_t.

Зазвичай значення шлюзів і прихованого стану ініціалізуються за допомогою нормального розподілу [19]. Простий спосіб зменшити проблему згасання градієнта — встановити $f_t = 1$.

Таким чином, LSTM-комірка успішно відфільтровує несуттєву інформацію на кожному часовому кроці та зберігає лише важливу.

LSTM є найпопулярнішою та найефективнішою формою рекурентної нейронної мережі, яка покращила показники якості багатьох моделей обробки природної мови (NLP) [19]. Однак мережа LSTM є обчислювально дорогою, що обмежує можливості її застосування на великих текстових наборах даних (розміром понад мільярд елементів). У 2017 році було розроблено нейронну мережу на основі механізму самоуваги (self-attention) [17], яка може працювати так само добре, як LSTM, майже на всіх NLP-наборах даних, але значно швидше.

Подібно до моделі LSTM, для вхідних подань $X = (x_1,...,x_N)$, математична модель самоуваги формує "запит" (Q) як $Q=XW^Q$, "ключ" (K) як $K=XW^K$, та "значення" (V) як $V=XW^V$. Тут W^Q, W^K і W^V — матриці моделі, коефіцієнти яких зазнають тренування.

Запит Q і ключ K описують відносну важливість кожного подання x_n у вхідних даних. Потім запит Q і ключ K взаємодіють зі значенням V, щоб отримати вихід моделі самоуваги:

$$Attention(X) = Softmax(\frac{QK^T}{\sqrt{d_k}})V, \quad (1.12)$$

Спрощуючи, можна сказати, що замість обчислення градієнтів уздовж часових кроків, як у RNN, математична модель самоуваги безпосередньо оцінює важливість кожного вхідного подання x_i у заданому контексті подань X. Потім, на основі цієї оцінки важливості, інформація, яку несуть вхідні подання, перетворюється та передається далі за допомогою операції уваги.

З моменту представлення цієї архітектури, вона стала де-факто стандартом індустрії — так Google використовує BERT модель у своїй пошуковій системі, а GPT моделі компанії OpenAI постійно з'являються в

заголовках новин [19]. Ця архітектура використовується в таких додатках як GitHub Copilot чи для модерації коменарій на платформі Reddit. І останні досягнення в обробці природної мови пов'язані з моделями саме на основі цієї архітектури.

Ця архітектура складається з повнозв’язаних шарів та механізму самоуваги, які утворюють шари Трансформеру, які, у свою чергу, утворюють кодувальну та декодувальну частини. Це перша архітектура, яка дозволяє працювати з послідовностями, лише спираючись на концепт механізму самоуваги, без рекурсивності чи згорток.

В даній дисертаційній роботі розглядаються саме моделі на основі архітектури типу “Трансформер”.

1.3 Аналіз сучасного стану методів обробки природної мови щодо текстів та текстових документів

1.3.1 Передобробка тексту для подальшого опрацювання

Перед застосуванням моделей для обробки тексту, необхідно провести передобробку даних, що включає кілька етапів. Основними завданнями передобробки тексту є його очищення та приведення до уніфікованого формату. Цей процес може включати токенізацію, нормалізацію, лематизацію або стемінг, видалення стоп-слів, усунення пунктуації та спеціальних символів. Моделі на основі архітектури типу “Трансформер” не потребують видалення стоп-слів чи спеціальної нормалізації даних [19]. Але токенізація є ключовим етапом в сучасних моделях. Наразі існує три популярних метода токенізації:

1. *Токенізація за словами*: розділяє текст на окремі слова. Це один із перших запропонованих типів токенізації, але він не є ефективним для мов, де слова зливаються разом або мають складні морфологічні форми. Крім того, цей метод передбачає використання спеціального токену для слів, яких немає

в словнику моделі, і тому всі слова, які не увійшли до словника, матимуть однаковий токен, що шкодить моделюванню подань таких слів.

2. *Токенізація за символами*: поділяє текст на окремі символи. Цей метод зазвичай використовується в ієрогліфічних мовах, таких як китайська, де кожен символ може відповідати окремому значенню.

3. *Токенізація підслів (subword tokenization)*: цей метод використовується у сучасних мовних моделях (наприклад, BPE — Byte-Pair Encoding, WordPiece). Цей метод передбачає розбиття слів на менші одиниці — підслова. Таким чином метод вирішує проблему невідомих слів або слів, які зустрічаються рідко, оскільки навіть нові чи складні слова можна представити як комбінацію підслівних токенів. Наприклад, у слові “рекурентний” токенами можуть бути “рекурент”, “ний”.

В даній дисертаційній роботі розглядається саме метод *Токенізація підслів* для токенізації.

1.3.2 Тренування моделей подань токенів

Найпоширенішим методом попереднього тренування моделей-енкодерів для подання токенів є стратегія “маскованого” тренування, вперше представлена авторами BERT [21]. Ця стратегія полягає в тому, що у вхідному тексті випадковим чином маскують 15% слів. Після цього модель отримує на вхід спотворений таким чином текст. Наступним кроком вхідні подання токенів $(x_1,\ldots,x_N)$, перетворюються моделлю у вихідні подання $(y_1,\ldots,y_N)$. Далі за допомогою лінійного шару подання замаскованих токенів перетворюються, що дає змогу передбачити ймовірності їх відповідних класів на основі контексту. Втрати обчислюються як крос-ентропія між передбаченими ймовірностями та реальними токенами, що стимулює модель навчитися враховувати контекст при обробці тексту. У цьому процесі модель вчиться “заповнювати пропуски” правильними словами для реконструкції тексту.

Традиційно в моделях обробки природної мови використовувалися два популярні підходи до обробки регістра. Перший підхід передбачає перетворення всього тексту у нижній регістр під час токенізації, що спрощує процес моделювання, але призводить до втрати важливої інформації про регістр. Так, дослідження з завданням розпізнавання іменованих сутностей (NER) показали видатні результати, що досягають рівня людини на текстах із правильним форматуванням, зокрема з правильною пунктуацією та використанням регістра. Автори роботи [22] демонструють, що правильне форматування тексту покращує якість передбачень в задачі NER майже на 4% за метрикою F1-score.

Альтернативно, другий підхід використовує словник, який зберігає оригінальний регістр слів. Проте цей метод має певні виклики, зокрема у випадках, коли слова повністю написані великими літерами: під час токенізації такі слова можуть бути розділені на кілька токенів, що може спотворити їхнє значення. Крім того, одне й те саме слово, написане в різних регістрах буде мати різні токени в словнику моделі. Це може ускладнити узагальнення моделі, оскільки вона сприйматиме варіації регістру як окремі лексичні одиниці, що призводить до фрагментації подань і збільшення розміру словника. Отже, існує потреба в більш гнучкому підході, який дозволив би ефективно враховувати інформацію про регістр, усуваючи недоліки традиційних методів.

Оскільки ефективне подання тексту є критично важливим для багатьох задач обробки природної мови, зокрема при обробці неструктурованих чи напівструктурованими даних, постає питання про адаптацію підходів до токенізації з урахування додаткових факторів, таких як регістр та стилі. Це особливо актуально у сценаріях, коли тексту притаманні візуальні ознаки. Традиційно для розв’язання таких задач текст поєднується з іншими модальностями, як-от зображення або структурні особливості документа.

Одним із ключових напрямів у цій сфері є візуальне розуміння документів (з англ. VDU – Visual Document Understanding), яке спрямоване на

аналіз візуально насичених документів, зокрема відсканованих сторінок. Часто для обробки таких документів застосовують методи розпізнавання тексту з картинки, а потім використовують текстові та позиційні (геометричне розташування елементів на сторінці) характеристики, а також комп'ютерний зір. Це дозволяє моделі отримати доступ до інформації, яка важлива для розуміння тексту людиною.

Інтеграція візуальної та стильової інформації в моделі на основі архітектури "Трансформер" для покращення обробки природномовних документів стала предметом багатьох досліджень останнім часом. Було запропоновано різні підходи до цієї проблеми з різним ступенем успішності. Моделі на основі архітектури типу "Трансформер" широко застосовуються для обробки візуально насичених документів [13, 14]. Ці моделі відрізняються методами попереднього тренування та імплементації механізму самоуваги, який дозволяє поєднувати ознаки з різних модальностей (текст, картинка, позиційні характеристики). Традиційною методикою тренування текстової модальності є запропонована в [21] стратегія маскованого мовного моделювання.

У дослідженні LayoutLMv3 [23] запропоновано техніку некерованого попереднього навчання, яка використовує завдання маскованого мовного моделювання для навчання двонаправлених подань токенів у текстовій модальності. Як додаткове завдання попереднього тренування використовується вирівнювання слів і фрагментів зображення для навчання міжмодального вирівнювання шляхом прогнозування, чи є відповідний фрагмент зображення текстового слова замаскованим. Ця модель продемонструвала найкращі результати як у тексто-орієнтованих, так і в зображення-орієнтованих завданнях обробки візуально насичених відсканованих документів з допомогою штучного інтелекту.

Аналогічно автори DocFormer [24] представили мультимодальну архітектуру на основі трансформера, розроблену для обробки візуально насичених документів. Представлена архітектура використовує текстові,

візуальні та просторові характеристики, комбінуючи їх за допомогою додаткового шару мультимодальної самоуваги.

Text-Image-Layout Transformer (TILT) [25] також вирішує проблему обробки документів поза використання суто текстової інформації. TILT одночасно вивчає інформацію про розташування текста, його візуальні особливості та текстову семантику. Інформація про розташування представлена як зміщення уваги та доповнена контекстуалізованою візуальною інформацією. Основою моделі є попередньо натренований “Трансформер” з архітектурою енкодер-декодер.

LamBERT [26] модифікує архітектуру енкодера “Трансформера”, використовуючи інформацію про розташування тексту на сторінці, отриману з OCR-системи, без необхідності повторного навчання мовної семантики з нуля. Модель лише доповнює вхідні дані координатами границь токенів без використання зображення.

1.3.3 Тренування моделей подань текстів

Навчання подань речень має важливе значення в різних задачах обробки природної мови (NLP), включаючи семантичний аналіз, машинний переклад і відповіді на запитання [17, 18]. Готові моделі подань речень з моделей класу BERT [21] часто не є конкурентними навіть в порівнянні з простими базовими моделями, такими як усереднення векторів GloVe [30] у задачах семантичної текстової подібності [31]. Це обмеження випливає з процедури навчання BERT, яке засновуються на парадигмі масковаго мовного моделювання. Ця стратегія тренування дозволяє моделі досягти видатних результатів в задачі подання токенів. Але для отримання оцінок подібності для пар речень обидва речення повинні оброблятися одночасно.

У зв’язку з цим були запропоновані спеціальні підходи для навчання подань речень і розроблено такі моделі як Universal Sentence Encoder [32] та SentenceBERT [31]. Ці моделі тренували при концепції керованого навчання,

використовуючи великі набори даних, марковані людиною, щоб покращити здатність моделей генерувати змістовні подання на рівні речень. Але необхідність мати великий обсяг даних, перевірених та маркованих людиною, обмежує використання цього підходу.

Для подолання проблеми в необхідності мати великі обсяги анотованих людиною даних були представлені підходи некерованого навчання векторних подань текстів. Зокрема з'явилося кілька підходів, які намагаються адаптувати ідею з самоконтрастним навчанням, яка набула поширення в комп'ютерному зорі. Ця ідея полягає у використанні різних “поглядів” на один і той самий зразок для побудови позитивних пар і випадкові зразки з тренувального набора — для негативних пар. Ефективна побудова цих різних “поглядів” на один і той самий текст є основним фокусом дослідників. Було запропоновано кілька підходів: у “Contrastive Tension” [33] автори пропонують тренувати два незалежних кодувальника; у “SimCSE” [34] автори використовують різні виключення (з англ. dropout) для побудови позитивних прикладів, що походять з одного тексту; “TSDAE” [35] будує позитивні приклади, додаючи до текстових даних певний шум (наприклад, видалення, перестановка, заміна слів).

Очевидним недоліком перших двох методів некерованого навчання, перелічених вище, є однакова довжина, притаманна позитивним парам, через їхню побудову з однакових речень. Це може призвести до того, що модель буде схильна створювати подання текстів однакової довжини близькими одне до одного, навіть якщо їх семантичний сенс різний. Для методу “TSDAE” [35] недоліком є велике перекриття слів в позитивних парах. Крім того, зашумлення даних потенційно може спотворити суть тексту, роблячи його граматично некоректним і важким для обробки моделлю.

Альтернативно були представлені методи, які використовують проміжки тексту, вибрані з одного документу для створення позитивних пар. Так, у “DeCLUTR” [36] позитивні пари вибираються як фрагменти тексту

довжиною 32-512 токенів, що не перекриваються, з одного документа. Фрагменти тексту вибираються випадковим чином.

1.3.4 Подання фраз у контексті

Підходи для тренування подань речень та текстів не розраховані для моделювання більш коротких лексичних одиниць, а саме фраз.

Подання на рівні фраз відіграють вирішальну роль у великому спектрі завдань, включаючи виявлення парафраз, відповіді на запитання та моделювання тем. Автори “PhraseBERT” [37] дотреновують BERT за допомогою синтетично згенерованих парафраз для виявлення лексично відмінних текстів. Але цей підхід покладається на синтетичні дані і буде обмежений знаннями, які містяться у великій мовній моделі, яку використовували для генерації даних.

В підході “McPhraSy” [38] автори пропонують використовувати 2-шаровий MLP, який поєднує подання, отримані від двох окремих попередньо навчених моделей. Так, основою виступають подання фрази, отримане з “PhraseBERT” [37] моделі, і подання маски на місці цієї ж фрази з випадкового речення, отриманого зі “SpanBERT” [39] моделі. На основі цих двох подань генерується остаточне контекстно-освічене подання фрази. Цей підхід демонструє ефективні результати та покладається лише на реальні тексти, не використовуючи великих мовних моделей для створення даних. Але має недоліки надлишкового використання обчислювальних ресурсів, бо використовує дві моделі.

Було створено еталони подібності речень [40, 41], щоб надати можливість порівнювати здатність моделей генерувати подання семантичної суті тексту. Ці набори даних представляє собою текстові пари, оцінені людьми щодо їх подібності.

Для фраз також було створено еталони подібності, серед яких загальні “PPDB” [42], “PIC” [43] та специфічні для конкретних доменів, такі як

“Patents Phrase to Phrase Semantic Matching Dataset” [44], “Amazon Reviews Dataset” [38]. Але хоча такий підхід з використанням текстових пар є загальноприйнятим для порівняння подань речень, використання аналогічної методики для фраз має суттєві недоліки. Дійсно, фрази можуть змінювати своє значення в залежності від контексту, і їхня подібність може бути неправильно інтерпретована без врахування цього контексту. А всі перераховані еталони подібності фраз мають спільний недолік щодо відсутності контексту. Все це дає підстави стверджувати, що доцільним є впровадження еталону подібності фраз з контекстними реченнями для кожної фрази.

1.3.5 Крос-лінгвістична дистиляція векторних подань

Велика кількість паралельних текстових фрагментів, доступних у таких наборах даних, як “OpenSubtitles” [45], “wikimatrix” [46] та інші, була зібрана та зроблена публічно доступною.

Було запропоновано низку методологій для навчання моделей багатомовного подання речень, серед яких можна виділити дві основні групи підходів для тренування моделей текстових подань, що використовують пари перекладених текстів.

Перший підхід полягає в навчанні моделі з єдиним енкодером за допомогою функції втрат для ранжування перекладів із використанням негативних прикладів у межах одного батчу [47]. Це завдання тренування полягає в узгодженні подань для джерельної та цільової мов у спільному векторному просторі. Моделі, навчені за цим підходом, вчаться проектувати подання одного й того самого речення різними мовами ближче один до одного, водночас розташовуючи подання інших речень, які не є прямими перекладами, далі одне від одного.

Однак, як було показано в численних дослідженнях [48], хоча стратегія використання негативних прикладів у межах одного батчу є досить

ефективною, вона вимагає значних розмірів батчу для досягнення оптимальної продуктивності, а більші розміри батчу є обчислювально інтенсивними. Другий недолік такої стратегії навчання полягає в тому, що моделі, навчені таким чином, часто стикаються з труднощами в оцінюванні подібності речень, які не є прямими перекладами одне одного [49].

Другий підхід передбачає використання моделі-вчителя, яка добре справляється з поданням речень мовою з великим обсягом ресурсів. Потім модель-студент навчається імітувати подання, згенеровані моделлю-вчителем для тексту цільової мови, і створювати подання, які тісно узгоджуються з поданнями моделі-вчителя у векторному просторі [49]. Цей підхід зменшує вплив розміру батчу на процес навчання, забезпечуючи, що якість подань, які вивчає модель-студент, залежить головним чином від можливостей моделі-вчителя, розміру датасету, а також його релевантності до текстів, які модель-студент буде опрацьовувати під час використання моделі.

Крім того, моделі, навчені для конкретної мови, зазвичай перевершують багатомовні моделі такого ж розміру [50]. Наприклад, у “LASER3” [50] дослідники пропонують використовувати модель, орієнтовану на певну мовну сім'ю, щоб отримати переваги від збільшення кількості зразків і перенести додаткові знання з інших подібних мов.

Іншим питанням, яке потребує подальшого дослідження, є функція втрат, яку доцільно використовувати для дистиляції текстових подань.

У роботі [49] автори пропонують використовувати середньоквадратичну похибку (MSE, з англ. Mean Squared Error) як функцію втрат. MSE широко застосовується як функція втрат у завданнях регресії і вимірює середню квадратичну різницю між передбаченими та фактичними значеннями. У контексті навчання текстових подань MSE порівнює векторні подання, створені моделлю-студентом, з тими, які генерує модель-вчитель, намагаючись мінімізувати розбіжності між цими двома наборами векторів.

Однак навчання з використанням функції втрат MSE може призводити до нестабільного навчання та неочікуваних результатів, особливо коли

векторні подання моделі-вчителя нормалізовані, що призводить до маленьких значень MSE [48]. Крім того, при використанні для вимірювання подібності зазвичай використовується функція косинусної подібності.

Косинусна подібність вимірює косинус кута між двома векторами і особливо підходить для порівняння подібності між векторами незалежно від їхніх величин. У контексті текстових векторних подань косинусна подібність визначає ступінь подібності між двома текстовими поданнями на основі напряму їхніх векторів у векторному просторі, а не їх абсолютних відстаней. Таким чином, вона є більш стійкою для відображення семантичної подібності між текстовими поданнями.

Крім цього, не було проведено дослідження як вибір моделі-вчителя впливає на якість текстових подань, вивчених моделлю-студентом, а також чинників, що позитивно або негативно на це впливають.

1.3.6 Генерація тексту

Генерація тексту є ключовим елементом в багатьох задачах обробки природної мови (NLP), включаючи відповіді на питання, машинний переклад, сумаризація тексту та створення діалогових систем.

Попереднє навчання для генерації тексту зазвичай передбачає авто-регресивний підхід, при якому модель навчають передбачати наступний токен на основі попередніх. Це може реалізовуватися як послідовне прогнозування токенів з урахуванням попереднього контексту (як у GPT [51]), або за допомогою моделей енкодер-декодерів, які генерують текст, враховуючи всю вхідну послідовність (такі моделі як BART [52] чи T5 [53]).

У процесі тренування модель оптимізує ймовірність правильної послідовності токенів, використовуючи функцію втрат на основі крос-ентропії. Такий підхід дозволяє генерувати зв’язний текст, що зберігає логічну послідовність і відповідає заданому стилю або тематиці.

Одним із важливих напрямів застосування таких моделей є сумаризація тексту (або текстова абстракція) — процес автоматичного створення скороченої версії тексту, яка зберігає найважливішу інформацію. Головна мета сумаризації полягає у зменшенні обсягу тексту без втрати змісту.

Залежно від методології та типу отриманого результату, сумаризацію тексту поділяють на кілька основних видів:

Екстрактивна сумаризація передбачає вибір найважливіших речень або фрагментів з вхідного тексту без змінення їхньої структури. Цей підхід спирається на аналізі значущості кожного речення з точки зору його інформаційної ваги. Традиційно для визначення ключових речень використовували такі алгоритми оцінки важливості речень, як PageRank або TF-IDF. Сучасні методи екстрактивної сумаризації включають визначення ключових речень з допомогою подань BERT моделі [54], або спираються на визначені ключових фраз [55]. Основними перевагами екстрактивної сумаризації є простота реалізації, швидкість обробки та мінімальний ризик “галюцинацій” моделі. Де під “галюцинаціями” мається на увазі, що модель генерує інформацію, якої немає у вхідному тексті, або створює некоректні твердження.

З іншого боку, екстрактивна сумаризація має кілька істотних недоліків, які випливають з дослівного копіювання речень документа. Так, оскільки цей метод працює виключно з наявними реченнями, він часто генерує тексти низької якості з точки зору стилю та зв’язності: виділені фрагменти можуть бути недостатньо узгодженими, створюючи уривчасту структуру. Також модель може включати повтори або несуттєві деталі, які знижують компактність і зрозумілість анотації. Крім того, екстрактивна сумаризація не здатна виконувати глибоке узагальнення, що обмежує її ефективність для текстів із високим рівнем абстракції чи складною логічною структурою.

Абстрактивна сумаризація передбачає генерацію нового тексту, який є узагальненням вхідного, використовуючи нові формулювання та синтаксичні конструкції. З поширенням нейронних мереж в поєднанні зі зростаючою

популярністю попередньо натренованих моделей і трансферного навчання цей метод став домінуючим. Провідні моделі використовують механізм самоуваги в блоках “Трансформер” та енкодер-декодер архітектуру, поєднуючи механізм копіювання та уваги і багатокритеріальні стратегії тренування [52, 53]. Так, автори PEGASUS [56] в процесі попереднього тренування використовують стратегію генерації ключових речень. Архітектура BART [52] передбачає спотворення вхідного тексту шляхом маскування токенів, перестановки речень та інших методик зашумлення тексту, після чого модель навчають відновлювати оригінальний текст. Це дозволяє BART ефективно “розуміти” контекст і генерувати високоякісні абстрактивні анотації [52]. Водночас T5 (Text-to-Text Transfer Transformer) пропонує універсальний підхід до обробки природної мови, перетворюючи всі задачі в текстовий формат, що дозволяє виконувати сумаризацію як завдання трансформації тексту.

Такі підходи дозволяються отримати більш невимушений та вільний стиль тексту, але достовірність та мінімізація галюцинацій все ще є викликом.

Останнім часом, великі генеративні мовні моделі демонструють вражаючі багатозадачні можливості, успішно виконуючи широкий спектр завдань обробки природної мови, включаючи сумаризацію, переклад та відповіді на запитання. Однак їх практичне використання часто супроводжується значними витратами. Наприклад, пропрієтарні моделі, такі як ті, що розроблені OpenAI, є закритими і доступними лише через платні API, що може призводити до значних операційних витрат при тривалому використанні. Крім того, ці моделі можуть викликати занепокоєння щодо конфіденційності, особливо під час обробки чутливих даних, таких як резюме.

З іншого боку, великі моделі з відкритим кодом, такі як сімейство LLaMA 3, пропонують більшу гнучкість і контроль. Однак більш потужні їх варіанти вимагають значних обчислювальних ресурсів, включаючи потужні GPU або високопродуктивні обчислювальні кластери для забезпечення

ефективного часу виконання. Ці вимоги до ресурсів можуть стати бар’єром для доступності, особливо для менших організацій або дослідників, які не мають доступу до сучасного апаратного забезпечення.

1.3.7 Дистиляція моделей

Хоча моделі на основі архітектури типу “Трансформер” досягли видатних результатів в багатьох завданнях обробки природної мови, такі моделі часто мають сотні мільйонів чи навіть мільярдів параметрів [57]. Крім того, дослідження показують [58], що попередньо натреновані моделі більшого розміру не лише досягають кращих результатів у прикладних завданнях, але й здатні навчатися на меншій кількості даних для досягнення високих результатів на прикладних завданнях. Але через обмеженість обчислювальних ресурсів та вимог щодо швидкості обробки документів практичне використання таких великих моделей часто є не є можливим та доцільним.

Дистиляція — методика, яка широко використовується для “стиснення” попередньо натренованих моделей. Вона полягає в тому, що меншу модель-студента тренують імітувати передбачення більшої, добре натренованої моделі-вчителя.

У керованому навчанні для завдань генерації тексту або класифікації токенів модель тренують передбачати відповідний клас для кожного токена. У завданні класифікації токенів модель передбачає заздалегідь визначений клас, тоді як у завданні генерації тексту вона прогнозує токен, обираючи його зі свого словника, що також є завданням класифікації. Традиційною тренувальною стратегією у таких випадках є мінімізаці крос-ентропічної функції втрат між передбаченнями моделі та “золотою міткою” класа. Добре натренована модель призначатиме високу ймовірність правильному класу, тоді як ймовірності інших класів будуть наближатися до нуля. Але ці наближені до нуля вірогідності все ж не завжди будуть такими і

відображають ступінь невизначеності моделі, що може бути корисним для оцінки її впевненості у прогнозах.

Існує два загальних підхода до дистиляції моделей. Перший — безпосереднє використання передбачень моделі-вчителя. Цей підхід зазвичай використовується, коли немає доступу до внутрішніх прихованих станів моделі-вчителя (наприклад, при використанні пропрієтарних моделей на кшталт GPT-4 компанії OpenAI). Другий підхід передбачає наявність доступу до прихованих станів моделі-вчителя та уможливлює тренування зі стратегією мінімізації різниці між показники моделі-вчителя та моделі-студента. При такому сценарії тренування необхідно, щоб обидві моделі мали однаковий словник, бо використовується розподіл цільових вірогідностей моделі-вчителя (м’які передбачення).

Формально процес дистиляції можна визначити як проблему оптимізації функції розподілу передбачуваних моделлю-студентом вірогідностей щодо зменшення розбіжностей з фіксованою функцією розподілу вірогідностей моделі-вчителя.

Так, “DistilBERT” [59] використовує м’які передбачення моделі-вчителя для маскованого мовного моделювання. Модель-студента ініціалізують шляхом вибірки 1 прихованого шару з кожних 2 шарів моделі-вчителя. Автори [58] пропонують використовувати дистиляцію розподілу самоуваги, і демонструють, що використання станів останнього прихованого шару моделі-вчителя для дистиляції дозволяє досягти кращих результатів в порівнянні з дистиляцією кожного шару.

Не було досліджено залежність якості дистиляції (здатності моделі-студента імітувати передбачення моделі-вчителя) від кількість прикладів, використаних під час для дистиляції для конкретних завдань, таких як класифікація токенів.

1.4 Автоматизована обробка природномовних текстів у сфері управління персоналом

1.4.1 Підходи до сегментації та парсингу

Першим кроком в системі аналізу неструктурованих даних, таких як резюме та вакансії, зазвичай виступає сегментація та парсинг.

Під сегментацією розуміється розділення тексту на структурні компоненти, такі як заголовки та розділи. Під парсингом розуміетьс витягнення ключових сутностей, таких як дати, назви компаній, назви посад, ключові навчи тощо.

Так, автори [59] експериментували з сегментацією резюме та виділенням ключових сутностей, використовуючи окремі моделі для кожного розділу резюме. Для сегментації тексту кожен рядок класифікували незалежно, без урахування контексту документа.

У роботі [60] автори зосередилися на італійських резюме. Їхня методологія передбачала початкову сегментацію резюме, досягнуту за допомогою підходу на основі пошуку шаблонів із використанням мовного та країнозалежного словника ключових слів. Після цього вони розробили окремі моделі послідовного маркування для різних розділів резюме. Архітектура компонента для послідовного маркування була заснована на використанні BERT архітектури як базової основи з додатковим шаром класифікації.

Автори [61] використовують алгоритм Boolean Naive Bayes для класифікації абзаців резюме на основі слів, що містяться в абзаці.

У дослідженні [62] автори пропонують розглядати задачу парсингу резюме як ієрархічну проблему послідовного маркування, одночасно передбачаючи теги розділів і сутностей на основі або подань, згенерованих за допомогою моделі T5 з зафіксованими вагами, або подань моделі fasttext із додаванням створених вручну ознак.

Дослідження [63] є, наскільки нам відомо, єдиним, яке пропонує

використовувати багату інформацію, закодовану у форматуванні резюме, для покращення їх парсингу. Вони пропонують додати спеціальні створені вручну маркери, такі як ознаки форматування та маркери, засновані на частоті, отримані з розміченого корпусу, наприклад, ознаку “частота в заголовках резюме”, до вхідних даних моделі BERT. Однак напівавтоматичний процес ідентифікації маркерів і необхідність кластеризації на основі розмічених зразків вводять потенційні обмеження, такі як залежність від якості та репрезентативності розмічених даних, що може впливати на здатність підходу до узагальнення.

1.4.2 Нормалізація сутностей у сфері управління персоналом

Для ефективного порівняння сутностей необхідно застосовувати нормалізацію. Тому останнім часом зросла зацікавленість у застосуванні спеціальних енкодерів у сфері, пов'язаній із управлінням людськими ресурсами.

Автори [64] використовують великий розмічений вручну набір даних зі схожими назвами вакансій і навчають сіамську рекурентну нейронну мережу для проекції представлення назв вакансій у векторний простір.

У дослідженні [65] автори запропонували розглядати це завдання як проблему багатозначної текстової класифікації та використовувати F1-метрику для оцінки ефективності класифікатора.

Особлива увагу приділяють вивченню змістовних подань назв посад на основі навичок, які з'являються в одному й тому ж описі роботи [66, 67].

Автори [66] навчають сіамську мережу для вилучення та агрегування ознак кожного токена в назві вакансії, використовуючи пов’язані з нею навички як навчальний сигнал. Модель оптимізується для прогнозування наявності або відсутності навички, подібно до підходу, застосовуваного в skip-gram моделі.

У роботі [67] автори використовують агреговане представлення навичок,

створене версією Distributed Bag of Words алгоритму Doc2Vec, і навчають енкодер-модель відображати представлення назви вакансії у вектор Doc2Vec, який відповідає її навичкам.

У сфері управління персоналом оцінка вимог до робочих місць була провідною сферою дослідження ринку праці протягом останніх двадцяти років [9]. Але групування та нормалізація навичок є менш дослідженою в порівнянні з представленням назв посад. Здебільшого до задачі підходять або як до Extreme Multi Label Classification (XMLC), При цьому за класифікаційну одиницю в деяких дослідженнях беруть речення [68], а в деяких – всю посадову інструкцію [69]. Після чого навчають моделі оцінювати семантичну відповідність між назвою навички та реченням, в якому вона з'явилася [68, 69]. Проте такий підхід це обмежує здатність таких архітектур адаптуватися до інших сфер застосування, бо мають фіксовану кількість класів, які здатні опрацьовувати.

В якості альтернативи було запропоновано підхід до групування навичок на основі даних [70], де автори використовували графову мережу. Підхід, описаний у їхньому дослідженні, концептуалізує кожну навичку як вузол у мережі, а зв'язки або вершини між цими вузлами встановлюються на основі двох ключових чинників: спільної присутності навичок у парі та лексичної подібності між різними навичками. Цей метод пропонує детальний погляд на те, як навички пов'язані між собою та впливають одна на одну в професійному контексті. Однак дослідження зіткнулося зі значною проблемою з точки зору обчислювальних вимог. Графові мережеві моделі, особливо з великою кількістю вузлів і зв'язків, вимагають значних обчислювальних ресурсів. Це обмеження стало очевидним в обсязі дослідження, оскільки автори були змушені проаналізувати лише 10 554 навички в своїй моделі, що значно поступається безлічі варіацій навичок, які зустрічаються в реальному середовищі.

Найбільш перспективні останні підходи покладаються на синтетичні дані, згенеровані ВММ [71, 72], показуючи, що ВММ здатні генерувати

реалістичні дані для подальшого навчання менших моделей. Однак моделі, навчені на синтетичних даних, можуть мати труднощі з узагальненням непередбачуваних ситуацій або нових описів навичок, які суттєво відрізняються від синтетичних прикладів. Додаткові труднощі можуть виникнути при виявленні нових навичок, які BMM не бачила через історичну природу своїх даних під час навчання. Це може обмежити здатність моделі ефективно обробляти реальні варіації навичок.

1.4.3 Сумаризація резюме

В галузі управління людськими ресурсами сумаризація документів може сильно підвищити ефективність роботи рекрутерів, надаючи їм стислий огляд довгих документів, таких як резюме. У роботі [73] автори провели дослідження ефективності різних моделей сумаризації резюме на основі датасету, що містить 280 резюме. Анотації до цих резюме були створені вручну, а для тестування було використано 75 із них. У ході експериментів було протестовано кілька моделей, і встановлено, що модель BART-large демонструє найкращі результати.

Однак обробка резюме за допомогою моделей на зразок BART-large стикається з обмеженням довжини вхідного тексту. Якщо резюме перевищує цей ліміт моделі, його доводиться скорочувати перед подачею в модель, що створює ризик втрати важливої інформації. Вибір оптимальної стратегії такого скорочення є критичною задачею, оскільки від нього залежить якість анотації. Автори [74] застосовують скорочення секцій та пропонують видалення стоп-слів, якщо довжина секції перевершує задану максимальную довжину. Якщо після видалення стоп-слів довжина секції все ще перевищує встановлену максимальну межу, скорочення тексту здійснюється шляхом видалення слів, які входять до 30% найчастотніших у документі.

Однак такий підхід до скорочення резюме має суттєві недоліки: він може призводити до втрати важливої інформації, особливо якщо ключові

слова також є частотними, що негативно впливає на якість подальшої обробки. Тому необхідно розробити більш ефективний підхід до скорочення тексту, який би зберігав ключову інформацію та мінімізував втрати важливих даних.

Другим обмеженням дослідження [73] є те, що автори використовують написані вручну анотації для тренування моделі. Але створення такого датасету вимагає значних людських ресурсів. Як альтернативу, доцільно використовувати великі мовні моделі, такі як GPT-4, для автоматичної генерації анотацій. Це дозволило б суттєво розширити обсяг даних для тренування моделей. Крім того, дослідження попереднього некерованого тренування є важливим напрямком, оскільки таке тренування дозволило б оцінити потенціал автоматичного збагачення моделей знаннями зі сфери застосування без значних витрат на створення анотацій.

Особливої уваги заслуговує аналізі ключових фраз і підходів до стимуляції моделі виділяти їх та відображати в анотації. Адже саме такі фрази визначають релевантність і точність підсумкової інформації.

Водночас важливо дослідити закон спадної віддачі для навчання моделей на синтетичних даних і встановити поріг, після перетину якого подальша генерація тренувальних даних не є доцільною. Такий аналіз дозволить оптимізувати використання ресурсів та створити більш ефективні підходи до навчання моделей.

1.4.4 Підходи до зіставлення резюме та вакансій

Системи зіставлення резюме та вакансій моделюють відповідність між резюме та вакансією, що дозволяє знайти найбільш підходящих кандидатів на посаду. Резюме (або оголошення про вакансію) зазвичай є напівструктурованим документом, який перевищує обмеження в 512 токенів, що накладається багатьма архітектурами моделей на основі трансформерів, такими як BERT. Це обмеження на кількість токенів створює проблему для

ефективного захоплення повної семантичної інформації з таких довгих документів.

У роботі [75] автори беруть перші 512 токенів документа як вхідні дані для моделі і навчають модель сіамської мережі на датасеті, який включає більше 274 тисяч маркованих людиною пар резюме-вакансія. Процес навчання включає в себе оптимізацію цілей класифікації та регресії для покращення продуктивності моделі. Однак таке скорочення документів може призвести до неоптимальної продуктивності через втрату інформації. Крім того, отримання датасета подібного масштабу є дуже ресурсозатратним.

Зі схожими обмеженнями стикаються автори [76], які обмежуються в дослідженні лише останнім місцем роботи кандидата, повністю ігноруючи всі інші деталі, присутні в резюме. Згодом вони навчають модель для завдання класифікації послідовних пар, де позитивні пари будуються на основі комбінацій посадових обов'язків, які виконувала людина, тоді як негативні пари складаються з пар обов'язків, що походять з різних профілів людей. Під час використання навчена модель отримує дані як про вакансію, так і про останнє місце роботи кандидата, а класифікаційний бал використовується для оцінки їх подібності. Обмеження розгляду лише останнього місця роботи може серйозно ускладнити роботу моделі, оскільки вона не враховує вимоги до освіти та сертифікації. Це обмеження особливо відчутно для нещодавніх випускників, які, можливо, не мали професійної роботи, пов'язаної з їхнім фахом або кар'єрними прагненнями. Багато випускників під час навчання в університеті влаштовуються на неповний робочий день або на роботу, не пов'язану з навчанням, що може не відповідати їхній кваліфікації, навичкам чи потенціалу.

Існують підходи, які намагаються вирішити цю проблему шляхом подання речень або фрагментів тексту окремо з подальшею агрегацією подань.

У роботі [77] автори навчають змагальну мережу реконструювати вакансію за її векторним представленням, а резюме розбивають на три

частини — досвід, навички та інші фактори, які об'єднують і пропускають через шари MLP, щоб отримати остаточне представлення резюме. Після цього представлення резюме та вакансії разом з їхнім поелементним продуктом об'єднуються і використовуються для бінарного прогнозування придатності кандидата до роботи. Такі порівняння, однак, стикаються з квадратичною обчислювальною складністю, оскільки підхід вимагає оцінювати кожне резюме на відповідність кожній вакансії в наборі даних.

У нещодавно представленому ConFit [78] автори намагаються вирішити деякі з вищезгаданих недоліків і пропонують підхід з використанням аугментації для моделювання подань резюме та оголошень про роботу. Вони пропонують використовувати доповнення тексту розділів — застосовуючи як методи EDA (Easy Data Augmentation), так і перефразування на основі BMM для генерації позитивних зразків, а потім переходити до навчання моделі кодування з використанням контрастного навчання з негативними зразками з батчу. Однак їхній підхід спирається на початковий маркований набір даних пар резюме-оголошення про роботу і застосовує аугментацію лише для збільшення кількості зразків.

Таким чином, незважаючи на те, що існуючі підходи дозволяють досягти хороших результатів при використанні маркованих людиною пар вакансія-резюме, не було представлено підходів до некерованого навчання, які б дозволили ефективно використовувати немарковані дані та зменшити залежність від ручної праці з маркування прикладів. Крім того, традиційний підхід до постановки задачі зіставлення резюме та вакансій як до задачі бінарної класифікації призводить до значних обчислювальних витрат через квадратичну складність порівняння всіх можливих пар вакансія-резюме. Нарешті, наразі не існує підходів до некерованого навчання текстових подань, спеціально розроблених для сфери управління людськими ресурсами, що залишає значну прогалину у використанні специфічних знань для більш ефективного та результативного підбору кандидатів та вакансій.

Висновки за розділом 1

1. Повністю автоматизована операція аналізу резюме та зіставлення вимог до кандидата з даними резюме виключає присутність рекрутера на цій операції етапу відбору кандидатів, що дає змогу знизити суб'єктивність у процесі найму. Однак, сучасні розробки в цій галузі не дають достатнього ефекту в співвідношенні швидкості та якості результатів, що потребує подальшого удосконалення та розроблення нових технологій ШІ в цьому напрямку.

2. Встановлено, що автоматизований аналіз документів в сфері управління персоналом із застосуванням ШІ, в тому числі обробка резюме для витягнення ключової інформації, зіставлення вакансій та резюме є необхідним елементом підвищення ефективності рекрутингу і перспективним напрямком для подальшого удосконалення і розвитку.

3. Встановлено важливість англійської мови для поширення набутих знань щодо використання штучного інтелекту у рекрутингу.

4. Моделі на основі архітектури “Трансформер”, які інтегрують візуальні ознаки, досягли значного прогресу для покращення обробки природномовних документів, але вони не передають стильову інформацію безпосередньо в модель. Вони або повністю ігнорують візуальні ознаки, присутні в документах, або передають все зображення як ознаку, використовуючи комп’ютерний зір для його обробки, що значно збільшує обчислювальні витрати. Тому доцільно запропонувати модель, яка б обробляла стильову інформацію, витягнуту безпосередньо з документів без використання технології комп’ютерного зору.

5. Аналіз літературних джерел дозволяє обґрунтувати доцільність використання контексту під час моделювання подань фраз, що важливо для розроблення технології контекстно-освіченого подання фраз.

6. Аналіз літературних джерел показує доцільність удосконалення підходів щодо скорочення текстів для подальшої сумаризації. Відсутність

даних щодо впливу обсягу тренувальних зразків при автоматичній їх генерації (з використанням великих мовних моделей) на якість сумаризації документів у домені управління персоналом потребує дослідження в цьому напрямку. Потребують дослідження підходи некерованого попереднього тренування з використанням структури документів, а також функції втрат, які використовуються для попереднього тренування (зважена функція втрат).

7. Встановлено необхідність удосконалення крос-лінгвістичної дистиляції векторних подань для підвищення ефективності рекрутингу із застосуванням ШІ. В завданні крос-лінгвістичної дистиляції текстових подань доцільно дослідити вплив функції втрат (косинусної подібності та середньоквадратичної похибки) на якість вивчених моделлю-студентом текстових подань, а також дослідити чинники, які впливають на якість подань моделі-студента при виборі моделі-вчителя.

8. Встановлено доцільність дослідження впливу дистиляції на показники швидкості та якості етапів технології ШІ щодо аналізу резюме та зіставлення з вимогами вакансій.

2 МОДЕЛІ ТА МЕТОДИ ОБРОБКИ ПРИРОДНОМОВНИХ ТЕКСТІВ

2.1 Доповнене подання токенів для візуально насичених документів на прикладі резюме

Традиційні методи подання токенів не враховують стильові ознаки, притаманні візуально насиченим документам [1]. Але значну роль в розумінні документів та правильному їх сприйнятті людиною відіграють стилі документів і форматування тексту. Тому метод представлення токенів для візуального насичених документів потрібно удосконалити для поліпшення якості автоматичної обробки таких документів. Це доцільно зробити шляхом безпосередньої інтеграції параметрів про стильові ознаки, де додаткові дискретні ознаки векторизуються і передаються в архітектуру “Трансформер” разом із позиційними і токеновими поданнями.

2.1.1 Вибір характеристик стилів документів і форматування тексту

Існує три найважливіші характеристики, які значно покращують представлення мови в документі, багатому на візуальний контент [1]:

— Інформація про використання великих літер;

— Розмір шрифту;

— Стилі шрифту.

Інформація про використання великих літер. Очевидно, що використання великих літер значно полегшує розпізнавання меж речень, власних назв, а також може слугувати засобом виділення або акцентування певних термінів чи фраз [23]. Враховуючи це, використання великих літер доцільно поділити на три окремі категорії: повністю великі літери, містять деякі великі літери, і повністю малі літери.

Розмір шрифту. Розмір шрифту є ще однією важливою ознакою, яка надає ключові підказки для читачів щодо структури документа та

пріоритизації інформації. Так, заголовки часто виділяють більш великим шрифтом для зручнішої навігації в документі. Щоб надати моделі доступ до такої візуальної ієрархії тексту, доцільно обчислити медіанний розмір шрифту у всьому документі, який буде слугувати точкою відліку. Після цього всі розміри шрифтів можна класифікувати у три категорії: більші за медіанний розмір шрифту, медіанного розміру та менші за медіанний розмір. Це має допомогти моделі визначити варіації у розмірі тексту, які можуть передавати акценти або ієрархію.

Стилі шрифту. Останньою важливою ознакою є стилі шрифту. Використання різних стилів шрифту, таких як звичайний, жирний, курсивний та підкреслений, дозволяє диференціювати та акцентувати певний зміст у документі.

2.1.2 Кодування стилів документів і форматування тексту

Для оптимізації аналізу було обрано набір бінарних ознак, які можуть набувати значення “1”, якщо вони застосовні до відповідного токена, або “0”, якщо ні. До цих ознак належать: “малий шрифт”, “середній шрифт”, “великий шрифт”, “жирний” та “курсив або підкреслений”.

Щоб перетворити ці бінарні атрибути стилю у стислу числову репрезентацію, доцільно застосовувати техніку маніпуляції з бітами. Ця техніка створює унікальний ідентифікатор стилю, позначений як “style_id”, який об’єднує стилістичні нюанси тексту. “style_id” обчислюється за наступною формулою:

$$style\_id = 1 + \sum_{i=0}^{n} feature_i * 2^i \quad (2.1)$$

Тут кожна змінна $feature_i$ представляє один із згаданих бінарних атрибутів стилю, що дозволяє ефективно кодувати багатогранні стилістичні нюанси тексту для подальшого аналізу та обробки.

Схему передобробки даних в моделі, яку назвемо CapStyleBERT (з англійської Capitalization and Style Aware BERT) представлено на рисунку 2.1.

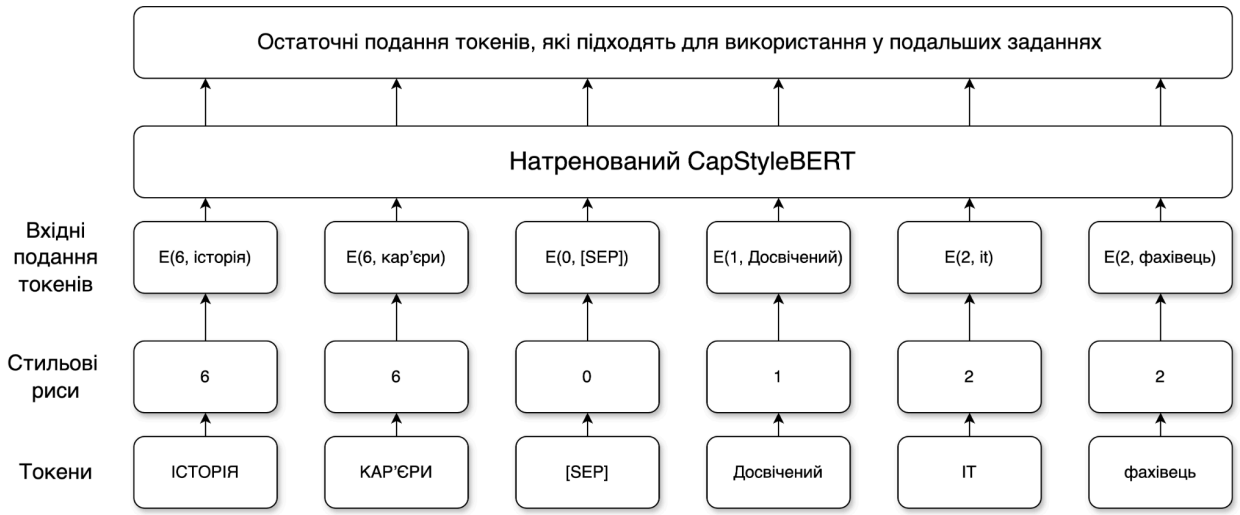


Рисунок 2.1 — Схема передобробки даних в CapStyleBERT, в якій інформація про стильові ознаки інтегрується в оригінальну архітектуру BERT

Можна показати, що запропонована CapStyleBERT модель дозволяє інтегрувати стильові ознаки, притаманні візуально насиченим документам, безпосередньо в модель на основі архітектури типу “Трансформер”.

2.2 Контекстно-освічене моделювання фраз на прикладі резюме та вакансій

2.2.1 Підхід до тренування подань назв посад, що базується на використання фраз навичок

Важливість ключових фраз в описі вакансій та резюме для розуміння посади не викликає сумнівів. Особливо це актуально у випадках, коли назва посади є загальною (Наприклад, посада “менеджер проектів” може стосуватися роботи в абсолютно різних галузях — від інформаційних технологій до будівництва або навіть організації подій, і, відповідно,

вимагати різних компетенцій). Різноманітність контекстів значно ускладнює автоматизований аналіз і нормалізацію назв посад, оскільки традиційні моделі зазвичай базуються на залежностях, сформованих під час тренування. Такі моделі під час використання отримують на вхід лише назви посад, тому їхні прогнози відображають тенденції, характерні для тренувальних даних, що може призводити до помилок у випадках з багатозначними назвами.

Тому доцільно розробити підхід до тренування подань назв посад, який би вирішив цю проблему шляхом визначення ключових фраз-навичок в тексті, а також введення спеціального токена “[SKILL]”, який слугував би для семантичного виділення та представлення кожної навички. Цей токен буде відповідати за представлення значення навички, що йде після нього. Для створення подання всіх навичок, які були в тексті, всі “[SKILL]” токени усереднюються, а отриманий вектор передається через лінійний шар для зменшення розмірності. Модель оптимізують шляхом мінімізації відстані між вектором назви посади та агрегованим поданням навичок.

Схематично підхід до тренування подань назв посад представлений на рисунку 2.2

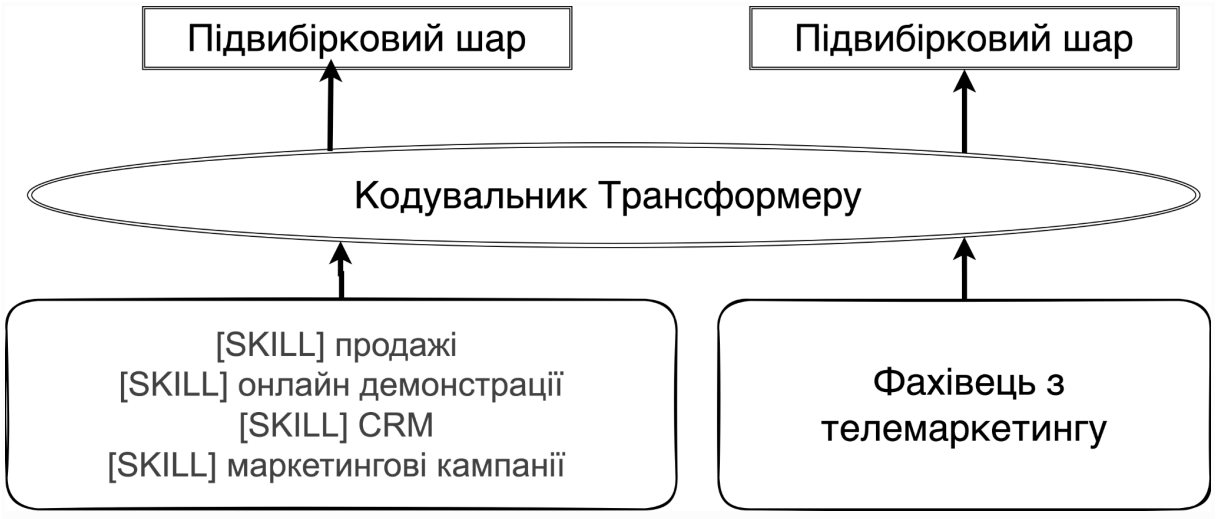


Рисунок 2.2 — Схема підходу до тренування подань назв посад

Таким чином, модель навчається наближувати подання назв посад та навичок з того ж опису роботи у векторному просторі. Це сприяє покращенню якості подань назв посад, оскільки дозволяє моделі краще

враховувати взаємозв’язок між посадою та необхідними компетенціями.

2.2.2 Архітектура представлення фраз у контексті з використанням спеціальних маркерів для виділення фраз для їх групування та нормалізації

Для більш загального сценарію моделювання фраз у контексті важливо враховувати структурні межі, які допомагають точніше інтерпретувати значення окремих частин тексту, а також розробити спеціальну архітектуру.

Основною гіпотезою є можливість використання спеціальних токенів для маркування меж фраз з метою підвищення точності моделювання подань фраз у контексті. Розробка моделі базувалася на припущенні можливості точно виділяти важливі фрази з тексту. Для спрощення для порівняння слід вибрати моделі розміру BERT-base.

Методи [69, 70] мають такі недоліки як обмеження дуже малим відсотком навичок, або від того, що розглядають ціле речення як мінімальну одиницю для класифікації, тоді як у реальних вакансіях це припущення справедливе лише для ~40 % речень вакансій (рисунок 2.3). Можна зробити припущення, модель також отримає додаткову корисну інформацію, якщо буде мати доступ до контексту більшого, ніж одне речення (а саме абзац).

Що стосується резюме, які здебільшого походять з PDF-документів, то проблема полягає у відновленні меж речень після процесу перетворення даних у зручний для подальшого опрацювання формат. Ця проблема є наслідком додавання додаткових символів нового рядка, що є притаманним обробці таких документів. Неправильне розділення речень за допомогою розділових знаків, які зазвичай позначають кінець речення, може призвести до втрати цінних навичок і забруднення контексту. Через це, при обробці секцій резюме може бути корисним зробити вибір на користь довших фрагментів тексту.

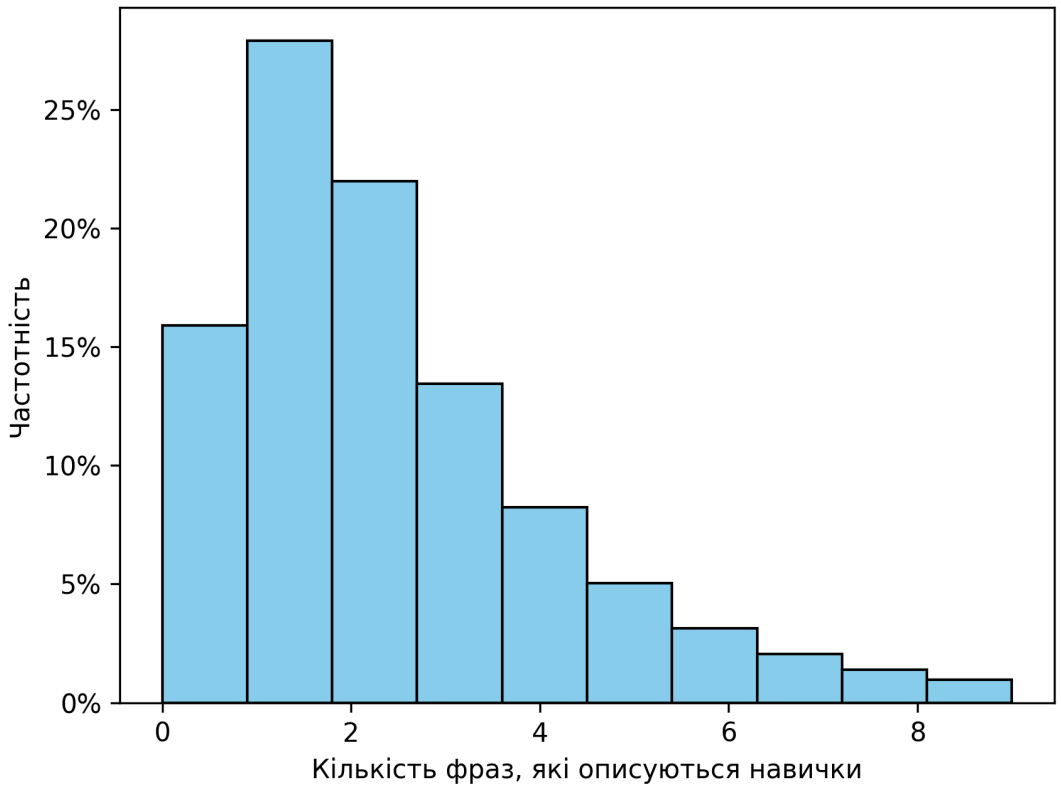


Рисунок 2.3 — Кількість ключових фраз, що описують навички, в одному абзаці у вакансіях

Метод навчання на основі багатоконтекстної подібності фраз [38] найкраще підходить для поставлених цілей. Дійсно, він враховує контекст не тільки під час навчання, але й під час використання, що дозволяє йому адаптуватися до фраз, які змінюють своє значення в різних контекстах. Також це надає моделі можливість безперешкодно працювати з новими фразами, які не були в наборі даних, на яких тренували модель. Однак цей метод має два недоліки: він використовує дві окремі базові моделі, що уповільнює його швидкість; а під час навчання автори пропонують налаштовувати лише кінцевий підвибірковий шар, ваги ж базових моделей залишаються незмінними. Це не дозволяє моделі повністю адаптуватися до нових доменів, відмінних від тих, які були в тренувальному наборі базових моделей. Крім того, як було показано на рисунку 2.3, в текстах, пов'язаних з галуззю управління персоналом (резюме/вакансії), в середньому 2,18 ключових фраз присутні в одному абзаці. Тому обчислення подань для кожної фрази шляхом

“прогону” того ж фрагмента тексту з іншою маскою через модель, а також обчислення подань для кожної фрази окремою моделлю є дуже витратним з точки зору обчислювальних вимог. З цього випливає, що логічним є зменшити потребу в обчислювальних ресурсах, вводячи нову стратегію додавання в текст спеціальних маркерів. Ці маркери, що відмічають початок і кінець ключових фраз, надають можливість отримати всі векторні подання лише за один “прогін” через модель.

Таким чином, метод багатоконтекстної подібності фраз [38] необхідно вдосконалити. Для досягнення цієї мети потрібно використати єдину модель, навчену за стратегією контекстно-орієнтованого вирівнювання фраз, де кожна фраза позначена своїми межами, а вектор, який відповідає токену лівої межі приймається за подання фрази. Удосконалений таким чином метод багатоконтекстної подібності фраз для подання фрази визначили як Phrase2Vec.

Для позначення початку і кінця фрази доцільно використати два нових токена, ініціалізованих з додаткових спеціальних токенів BERT. Всі ключові фрази, виявлені в абзаці, позначаються таким чином.

Якірною фразою *ph* слід визначити будь-яку обрану фразу. Для такої якірної фрази *ph*, що належить абзацу t_i, вибирається випадковий абзац t_j, який містить позитивну фразу *ph*+ – ту саму фразу, що є якірною для навчання в умовах повної відсутності маркованих людиною даних. Для напівкерованого навчання такою фразою слугує фраза, яка була визначена як фраза-синонім великою попередньо навченою моделлю.

Після цього фраза *ph* в t_i маскується за допомогою одного токена “[MASK]”, а фраза *ph*+ в t_j, яка слугує позитивним прикладом, залишається без змін.

Векторне подання токену, який позначає ліву межу фрази було використано як подання фрази.

Концепція контрастного навчання, особливо з негативними прикладами з того самого тренувального батчу даних, стала поширеною технікою в

дослідницькому співтоваристві для навчання репрезентативних текстових подань. Ця стратегія навчання дозволяє ефективно повторно використовувати обчислені подання під час навчання для кожної можливої пари в батчу даних. На основі цієї концепції використовується функція множинних негативних втрат при ранжуванні. Ця функція втрат працює, розглядаючи кожну пару подань (q_j, p_i) в межах заданого батчу. Розмір цього батчу визначається як k. Множинні від'ємні втрати при ранжуванні обчислюються для кожної з цих пар за формулою (2.2):

$$Loss_{q_j, p_j} = \quad - log_e \frac{e^{s(q_j, p_j)}}{\sum_{i=1}^{k} e^{s(q_j, p_i)}}, \tag{2.2}$$

де $s(q_j, p_j)$ — косинусна подібність векторів q_j і p_j; q_j — подання замаскованої фрази на j-ій позиції в батчі; p_j — подання не замаскованої фрази на j-ій позиції в батчі; e — основа натурального логарифму; log_e — натуральний логарифм; k — розмір батчу, $Loss_{q_j, p_j}$ — втрати щодо зазначених індексів для пари q_j,p_j

Далі модель навчають з метою мінімізації відстані між поданнями токену лівої межі фрази в позитивних парах. Цей токен має відображати значення конкретної фрази, що розглядається, в обох сценаріях: замаскованому (запит q_j) і не замаскованому (позитивний p_j). Одночасно відбувається підсилення розбіжності у поданнях не пов'язаних фраз, тим самим сприяючи вивченню моделлю якісних подань фраз.

Результати дослідження [48] підкреслюють переваги обчислення втрат у двонаправленому режимі (“всі негативні пари”) при тренуванні подань речень. Таких досліджень не проводилося для тренування подань фраз при стратегії маскованого зіставлення, тому доцільно перевірити вплив підходу “всі негативні пари” та порівняти з традиційним підходом до тренування.

Математично, розрахунок втрат для підходу “всі негативні пари” можна

виразити наступним чином:

$$Loss_{p_j, q_j} = -log_e \frac{e^{s(p_j,q_j)}}{\sum_{i=1}^{k} e^{s(p_j,q_i)} + \sum_{i=1, i\neq j}^{k} e^{s(q_j,q_i)} + \sum_{i=1, i\neq j}^{k} e^{s(p_j,p_i)}}, \quad (2.3)$$

де $s(q_j, p_j)$ — косинусна подібність векторів q_j і p_j; q_j — подання замаскованої фрази на j-ій позиції в батчі; p_j — подання не замаскованої фрази на j-ій позиції в батчі; e — основа натурального логарифму; log_e — натуральний логарифм; k — розмір батчу, $Loss_{p_j, q_j}$ — втрати щодо зазначених індексів для пари q_j,p_j

Мета мінімізації рівняння (2.3) полягає в тому, щоб змусити функцію подань створювати близькі у векторному просторі подання для токенів, що представляють однакові (або синонімічні) фрази. В той же час ця функція втрат забезпечує більше розділення для подань не пов’язаних між собою фраз в батчі. Розрахунок втрат при підході “всі негативні пари” відрізняється від класичного тим, що додатково враховується подібніть між замаскованими фразами та подібність між незамаскованими фразами в батчі. Такий підхід потенційно дозволить збільшити кількість негативних прикладів в батчі, що, як відомо [48], покращує якість векторних подань. Але, враховуючи асиметричну природу даних в батчі (замасковані та не замасковані фрази) використання такої стратегії тренування потребує подальшого дослідження та перевірки ефективності для тренування контекстно-освічених подань фраз.

На рисунку 2.4 представлена архітектура контекстно-освіченого подання фраз. На відміну від класичної архітектури моделі типу “Трансформер” [21], запропонована архітектура включає додаткові блоки, що дозволяють визначити межі фраз (блок “Перетворення Даних” на рисунку 2.4). Таким чином, запропонована архітектура є модифікацією класичної архітектури типу “Трансформер”.

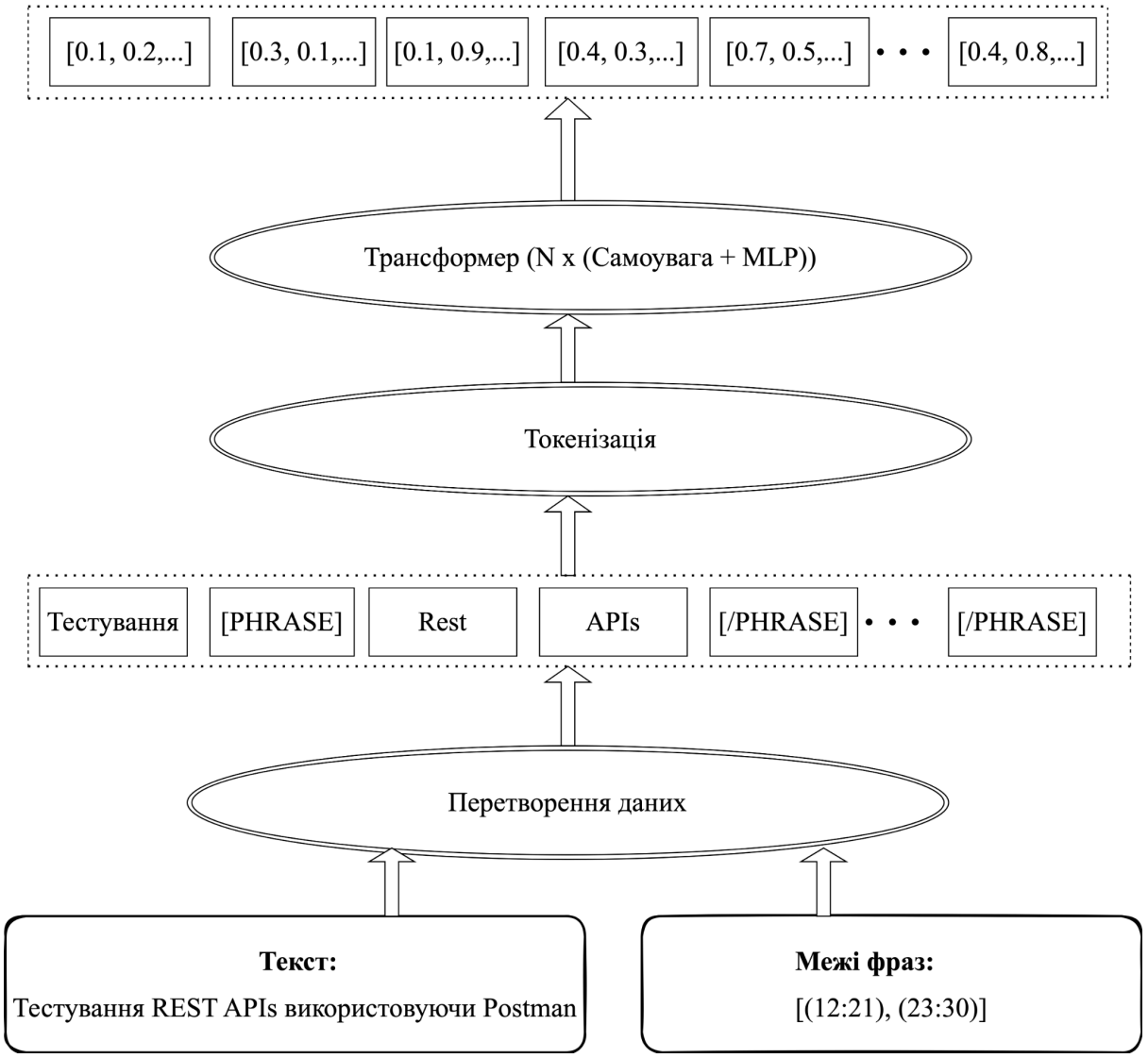


Рисунок 2.4 — Архітектура моделі контекстно-освічених подань фраз

Взаємозв’язок між перетворенням даних і архітектурою контекстно-освіченого подання фраз полягає в тому, що правильне визначення меж фраз є ключовим для ефективної роботи нейронної мережі (“Трансформеру”). Блок “Перетворення Даних” (рисунок 2.4) забезпечує уніфікацію вхідних даних шляхом маркування меж фраз, що дозволяє запропонованій архітектурі (рисунок 2.4) інтегрувати цю інформацію під час обчислення подань. Це, в свою чергу, сприяє точнішому моделюванню семантики фраз та покращенню якості їх подання у порівнянні з класичною архітектурою “Трансформер”, яка не має спеціалізованих механізмів для обробки фразових меж.

2.3 Метод некерованого навчання моделі векторних подань з використанням структури документів

Через обмеженість наявних створених людиною тренувальних даних для задачі моделювання векторних подань текстів для специфічних доменів та високу трудомісткість створення таких даних, важливо дослідити некеровані стратегії попереднього навчання моделей векторного подання текстів. В домені управління персоналом резюме, написані людьми, є цінним і природним набором даних для цієї мети, оскільки вони містять багатий структурований зміст і лінгвістичні нюанси з різних професійних сфер. Це робить їх ідеальним ресурсом для попереднього некерованого навчання моделей у задачах обробки природної мови, таких як семантична подібність текстів.

Базуючись на ідеї, представленій у “DeCLUTR” [36], де позитивні пари вибираються як фрагменти тексту, що не перекриваються, з одного документа, цей підхід можна покращити, використавши внутрішню структуру, притаманну резюме. Таку стратегію назвали HR Embedder. Вирішуючи проблему випадкового вибору текстових фрагментів, удосконалені стратегії вибору повинні враховувати структуровану природу резюме. Наприклад, секція “анотація”, яка покликана надавати загальний огляд професійного профілю особи, може бути поєднана з описом найостаннішої посади із секції про досвід роботи. Це дозволить використати природну відповідність між загальним професійним оглядом з анотації та конкретними робочими обов’язками, забезпечуючи семантичну узгодженість та контекстуальне багатство позитивних пар.

Аналогічно, у межах секції про досвід роботи часто присутні рядки, оформлені списковим чином. Такі рядки зазвичай описують аспекти роботи, роль і обов’язки, які виконував працівник, перебуваючи на конкретній посаді. Ці рядки мають спільний контекст однієї посади, описуючи різні аспекти виконуваних обов’язків. Наприклад, людина може написати “Чистка та пилососіння килимів” і “Прання; застилання ліжок” в одному описі роботи,

акцентуючи увагу на різних аспектах однієї виконуваної посади. Тому замість випадкового вибору фрагментів тексту логічно формувати позитивні пари шляхом інтелектуального вибору взаємодоповнюючих або послідовних завдань, які разом описують специфіку роботи. Таким чином, випадковий вибір таких завдань з одного опису роботи для формування позитивної пари змусить модель створювати нюансовані векторні подання текстів, які відображають сутність конкретних робочих функцій. Це призведе до покращення "розуміння" моделлю асоціацій між навичками та обов'язками в межах певних ролей і підсилить її здатність моделювати контекстуальну узгодженість у межах ролей.

Розділ "анотація", який зазвичай тісно пов'язаний із останньою посадою кандидата, у поєднанні з рядками, що описують реальні завдання з останнього досвіду роботи, доцільно використати в якості додаткового джерела для створення позитивних пар. Анотації часто охоплюють ключові навички, досягнення та обов'язки, подаючи стисле відображення основних сильних сторін і сфер експертизи кандидата. Коли рядки з анотації використовуються в якості позитивної пари з рядками з секції про останній досвід роботи, вони утворюють сильні, контекстуально узгоджені позитивні пари. Це повинно підсилити здатність моделі вловлювати основні компетенції, релевантні професійному профілю кандидата.

Цей підхід дозволяє моделі краще узагальнювати інформацію з резюме різних структур і стилів, створюючи більш глибоке розуміння того, як описуються ролі та навички. Завдяки включенню як завдань, орієнтованих на конкретні функції, так і описів високого рівня в контрастне навчання, модель може перетворювати тест у векторні подання, які не лише контекстуально узгоджені з конкретними функціями роботи, але й широко застосовні до різних сценаріїв зайнятості, що сприяє підвищенню точності підбору резюме щодо вакансій.

Крім того, вакансії, в яких прописуються вимоги до кандидатів теж можуть надати корисний контекст для формування позитивних пар. Вакансії

часто містять важливі деталі про обов'язки, кваліфікації та навички, необхідні для виконання конкретної ролі. Вакансії можна поділити на три основні сегменти, а саме: “вимоги” (опис необхідного досвіду, освіти, сертифікацій тощо), “роль” (майбутні обов’язки працівника), “опис компанії та переваг”. Після цього сегменти “вимоги” та “роль” доцільно використати в якості позитивної пари. Таке використання надасть моделі додаткового “розуміння” взаємозв’язків між вимогами щодо компетенцій та функціональними обов’язками.

У парадигмі контрастного навчання при стратегії множинного ранжування негативні приклади вибираються з того ж батчу. Зокрема, всі зразки в батчі, окрім позитивного для конкретного зразка, використовуються як негативні для нього. Це змушує модель вивчати дискримінативні ознаки та розрізняти семантично пов’язані та випадкові тексти.

Мірою подібності доцільно обрати косинусну подібність через її здатність вимірювати кутову відстань між векторними представленнями у високовимірних просторах незалежно від величини векторів. З огляду на результати досліджень інших науковців, функцію втрат доцільно обчислювати в обох напрямках.

$$Loss_{s_j, emp_j} = -\log_e \frac{e^{sim(s_j, emp_j)}}{\sum_{i=1}^{k} e^{sim(s_i, emp_j)} + \sum_{i \neq j}^{k} e^{sim(s_i, s_j)} + \sum_{i \neq j}^{k} e^{sim(emp_i, emp_j)}}, \quad (2.4)$$

де $sim(s_j, emp_j)$ — косинусна подібність векторів s_j та emp_j; s_j — векторне подання “анотації” витягнутої з резюме на j-ій позиції в батчі; emp_j — подання секції “досвід роботи”, витягнутої з резюме, яка знаходиться на j-ії позиції в батчі; e — база натурального логарифму; $\log_e$ — натуральний логарифм; k — розмір батчу; $Loss_{s_j, emp_j}$ — втрати стосовно конкретного індексу j в батчі.

$$Loss_{p1_j, p2_j} = -\log_e \frac{e^{s(p1_j, p2_j)}}{\sum_{i=1}^{k} e^{s(p1_i, p2_j)} + \sum_{i \neq j}^{k} e^{s(p1_i, p1_j)} + \sum_{i \neq}^{k} e^{s(p2_i, p2_j)}}, \quad (2.5)$$

де $s(p1_j, p2_j)$ — косинусна подібність векторів $p1_j$ і $p2_j$; $p1_j$ — векторне подання частини опису роботи (або розділ про досвід роботи в резюме, або розділ опису вимог у вакансії), який знаходиться на j-й позиції в батчі; $p2_j$ — векторне подання іншої частини опису позиції, що знаходиться на j-й позиції в батчі; e — основа натурального логарифма; log_e — натуральний логарифм; k — розмір батча; $Loss_{p1_j,p2_j}$ — втрати, пов’язані з указаним індексом j у батчі.

Остаточні втрати розраховуються за формулою:

$$Loss = \sum_{j=1}^{k} (Loss_{s_j,emp_j} + Loss_{p1_j,p2_j}), \quad (2.6)$$

де *Loss* — це остаточні втрати для батчу розміром k.

Під час навчання обчислюється та максимізується з допомогою функції втрат косинусна подібність між векторами пов’язаних текстових пар, водночас подання нерелевантних (випадкових) текстових пар відштовхується у векторному просторі. Таким чином, модель HR Embedder навчається ефективно порівнювати тексти у сфері управління персоналом.

2.4 Метод автоматичного створення датасету вакансія-резюме

Для тренування моделі зіставлення вакансій та резюме необхідний датасет із зіставленими парами “вакансія-резюме”.

Для отримання такого датасету з резюме шляхом сегментації витягували інформацію про останнє місце роботи. Далі цю інформацію за допомогою ВММ (GPT-4) перетворювали в опис вакансії, яку зіставляли з оригінальним резюме з видаленим з нього описом останнього місця роботи. Задана ВММ інструкція повинна бути структурована таким чином, щоб спрямувати модель на генерацію опису вакансії, що охоплює основні елементи, такі як обов'язки, необхідна кваліфікація та бажані навички. Видалення запису про останнє місце роботи з оригінального резюме гарантує, що створений опис вакансії не буде безпосередньо відображений у

резюме, імітуючи реальний сценарій, коли кандидат подає заявку на посаду, але його резюме не містить ідеального збігу з вакансією. Видаляючи запис про попереднє місце роботи, синтетична пара стає валідним навчальним прикладом для моделі, імітуючи реальне завдання зіставлення резюме з вакансіями.

Такий метод не лише усуває потребу у вручну анотованому наборі даних, але й забезпечує більшу гнучкість і масштабованість, дозволяючи навчати модель на різноманітних і актуальних прикладах. Таким чином стає можливим створення масштабного синтетичного датасету з парами вакансія-резюме.

Інші варіанти отримання великомасштабного анотованого датасету зі зіставленими парами вакансія-резюме в більшості реальних сценаріїв можуть бути недосяжними через трудомісткість і високу вартість створення таких наборів даних. Ручна анотація резюме та вакансій вимагає експертних знань у доменній області, що значно збільшує витрати. Більше того, динамічний і постійно змінюваний характер ринку праці, з постійною появою нових ролей і вимог, означає, що існуючі анотовані набори даних можуть швидко застаріти. Як наслідок, моделі, навчені на статичних, вручну створених даних, можуть мати труднощі з узагальненням нових описів вакансій і резюме.

2.5 Методи сумаризації документів у домені управління персоналом

2.5.1 Метод скорочення тексту з урахуванням структури документа та ключових фраз для подальшої сумаризації

На рисунку 2.5 представлено розподіл кількості токенів в резюме при використанні BART-large токенізатору на резюме, з яких були видалені “персональна інформація” та “рекомендації” секції.

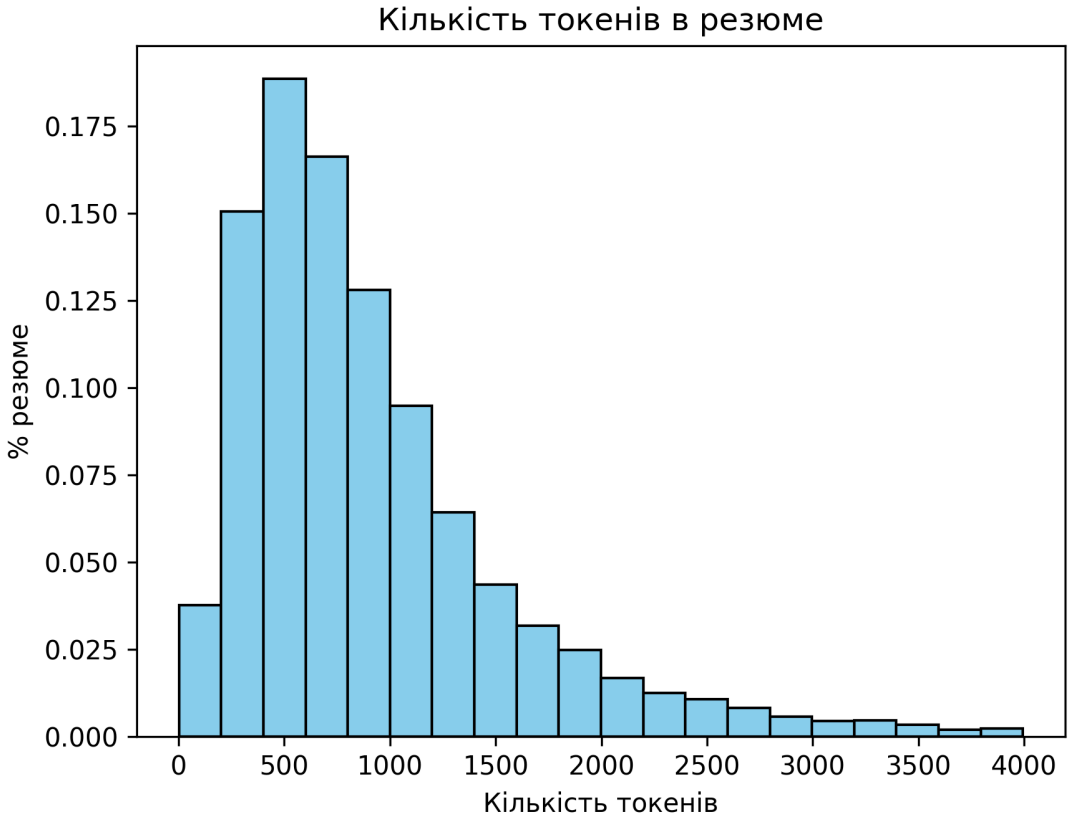


Рисунок 2.5 — Розподіл кількості токенів в резюме

Як видно з рисунку 2.6, 32% необроблених резюме перевершують довжину, яку спроможний обробляти BART-large, обмежений довжиною контексту в 1024 токени.

Для скорочення тексту, яке необхідне в випадку довгих документів (довжина секції перевищує встановлену максимальну межу), текст необхідно скорочувати з кінця, проходячи речення за реченням. З кожного речення виділяють і зберігають лише ключові фрази, які визначають спеціально навченою моделлю для виділення ключових фраз (парсинг) за алгоритмом, наданим на рисунку 2.6

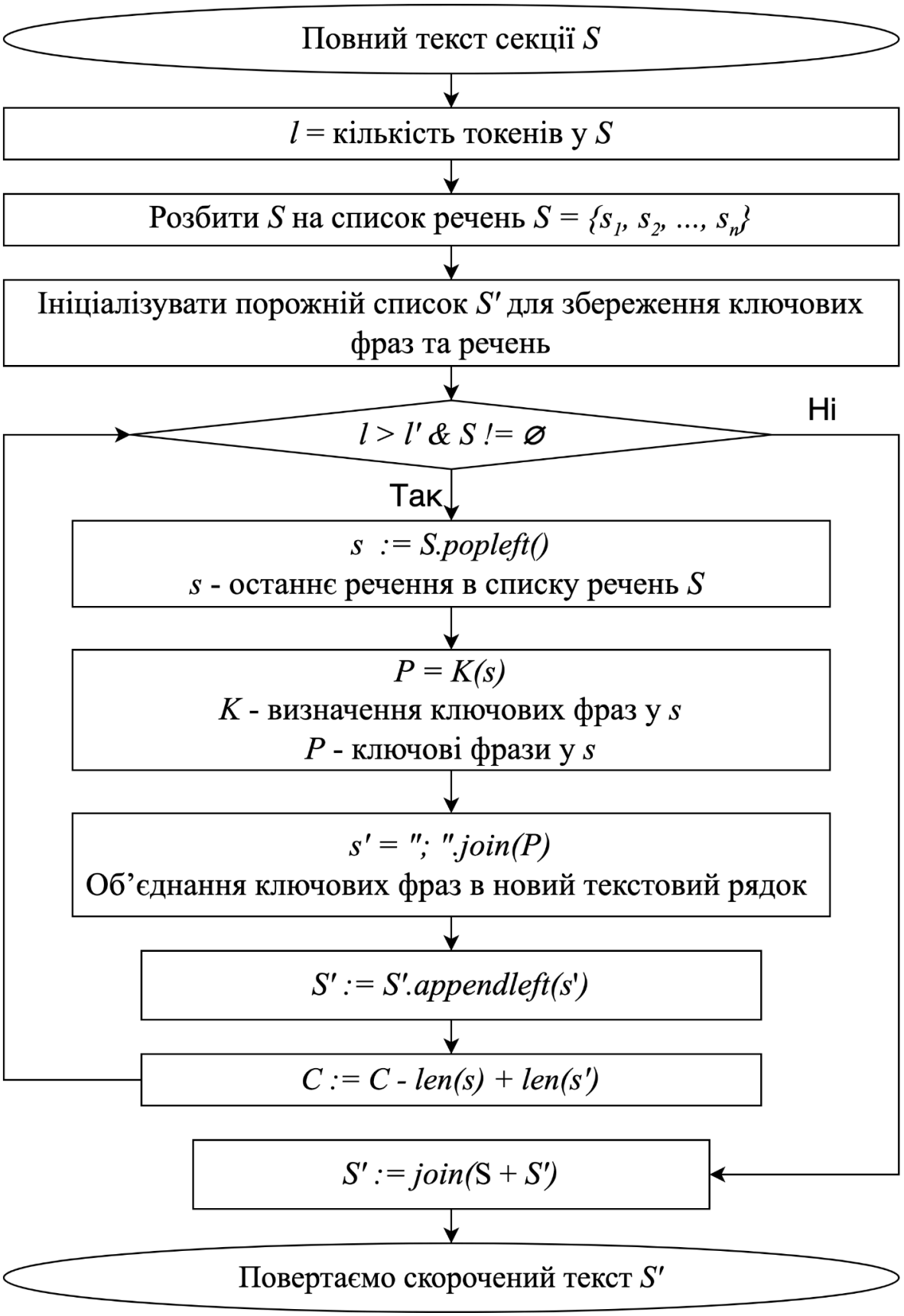


Рисунок 2.6 — Алгоритм зменшення обсягу тексту секції з використанням ключових слів

Текст резюме поділяли на секції як описано в 2.1. Якщо сумарна кількість токенів секцій (за винятком секцій “персональні дані” та “рекомендації”, які не використовуються для подальшої сумаризації) перевершує максимальну довжину, яку спроможна обробляти модель, то секції скорочували. Кількість токенів, до якої необхідно скоротити секцію розраховували динамічно з урахуванням її довжини та відсоткового внеску у загальну довжину резюме за наступною формулою:

$$l_j^{'} = l_j * \frac{M}{L}, \quad (2.7)$$

де L — загальна кількість токенів резюме; M — максимальна довжина, яку здатна обробляти модель; l_j — кількість токенів j-ої секції; $l_j^{'}$ — довжина, до якої необхідно скоротити j-у секцію.

Таке визначення кількості токенів секції дозволяє забезпечити збереження співвідношення довжин секцій резюме.

2.5.2 Некероване навчання з використанням структури документів для сумаризації

Використання ВММ, таких як GPT-4o для створення тренувальних прикладів, особливо в задачах генерації тексту, набула широкого використання, дозволяючи значно пришвидшити формування датасетів. Однак, така генерація, хоча й коштує менше, ніж коли це робить людина, все ще є дорогою, особливо при необхідності забезпечити великих обсяг даних. Тому дослідження методів некерованого навчання, які дозволяють зменшити залежність від генерації нових прикладів і підвищити ефективність використання наявних даних, є актуальним напрямом для зниження витрат та покращення результатів моделей.

Резюме можуть слугувати цінними даними для попереднього навчання для завдання сумаризації, оскільки приблизно 35% резюме містять короткий розділ “анотація” (summary). Проте, незважаючи на те, що більшість із них

пропонують багатий контекстуальний огляд професійного досвіду та кваліфікацій, ці анотації не мають уніфікованої структури, часто є надмірно багатослівними та включають загальні “м’які” навички або нерелевантні деталі. Навчання на таких анотаціях безпосередньо може призвести до того, що модель генеруватиме повторення та найпоширеніші “м’які” навички (як-от “працьовитість”, “робота в команді”). Однак, маючи наперед визначені ключові фрази, цю проблему можна вирішити за допомогою удосконаленого врахування ключових фраз у функції втрат. Зокрема, маючи попередньо ідентифіковані ключові фрази в тексті можливо підвищити вагомість для токенів, які формують ключові фрази під час обчислення втрат моделі. Це збільшує важливість токенів, пов’язаних із ключовими фразами, зменшуючи при цьому внесок у втрати загальних токенів, що сприяє створенню більш інформативного контенту.

Зважену функцію втрат з урахуванням ключових фраз можно описати наступною математичною моделлю:

$$Loss = \frac{1}{N}\sum_{i=1}^{N} w_i * l(y_i, \hat{y}_i), \quad (2.8)$$

де N — кількість токенів в тексті, що генерується; y_i — справжнє значення для *i*-го токена; $\hat{y}_i$ — прогнозоване значення *i*-го токена; l — базова функція втрат (Cross-Entropy Loss); w_i — вага для *i*-го токена (w_i = α, якщо токен є ключовим словом, w_i = 1 — в усіх інших випадках)

α = 3,4 для ключових фраз резюме [79].

2.6 Визначення часу ознайомлення рекрутером із резюме та анотаціями резюме

Для визначення часу ознайомлення з резюме та анотаціями резюме було проведено дослідження, для якого було залучено 253 рекрутери з

компаній, які користуються продуктами Daxtra Technologies. Рекрутери протягом 10 хвилин ознайомлювалися з документами (окремо з резюме, окремо з анотаціями резюме). Отримані дані (кількість резюме, переглянутих кожним рекрутером за 30 хвилин та кількість анотацій резюме, опрацьованих кожним рекрутером за 30 хвилин) піддавали подальшій обробці. Для дослідження було використано логування, яке дозволяє фіксувати відкриття резюме користувачем платформи (рекрутером). Респонденти (рекрутери) були повідомлені про проведення експерименту та надали свою згоду на участь.

За кінцеву мету дослідження було встановити наступне:

1. час, який витрачено на ознайомлення з повним резюме;

2. час, який витрачено на ознайомлення з анотацією резюме, сформованою за допомогою інформаційної технології.

Час на ознайомлення з одним документом, отриманий кожним респондентом (рекрутером) знаходили за формулами 2.9, 2.10:

$$\text{Тр}_i = \frac{1800}{\text{Кр}_i}, \quad (2.9)$$

де $Тр_i$ — час на ознайомлення з 1 резюме i-м рекрутером, $Кр_i$ — кількість резюме, з яким ознайомився i-ий рекрутер за 30 хвилин.

$$\text{Та}_i = \frac{1800}{\text{Ка}_i}, \quad (2.10)$$

де $Та_i$ — час на ознайомлення з 1 анотацією i-м рекрутером, $Ка_i$ — кількість анотацій, з якими ознайомився i-ий рекрутер за 30 хвилин

Середні значення за часом ознайомлення з резюме знаходили за формулою:

$$\text{Тр} = \frac{\sum_i^{\text{К}} Тр_i}{K_i}, \quad (2.11)$$

де $Тр_i$ — час на ознайомлення з 1 резюме i-м рекрутером, $К$ — кількість рекрутерів, які брали участь в експерименті.

Середнє значення часу ознайомлення з анотацією знаходили за формулою:

$$\text{Та} = \frac{\sum_{i}^{K} Та_i}{K}, \quad (2.12)$$

де $Та_i$ — час на ознайомлення з 1 анотацією i-м рекрутером, K — кількість рекрутерів, які брали участь в експерименті.

Висновки за розділом 2

1. Представлено новий метод безпосередньої інтеграції параметрів про стильові ознаки, де додаткові дискретні ознаки векторизуються і передаються в архітектуру “Трансформер” разом із позиційними і токеновими поданнями.

2. Запропоновано новий метод тренування подань назв посад, що базується на використання фраз навичок, які зазначені в описі роботи. Цей метод базується на введенні спеціального токена для виділення та представлення кожної навички у поєднанні з контрастним тренуванням з метою зіставлення усередненого подання навичок та назви посади з одного опису роботи.

3. Запропоновано новий метод некерованого навчання моделі HR Embedder з використанням структури документів. Цей метод являє собою модифікацію DeCLUTR, в якій позитивні пари для подальшого контрастного навчання вибирають наступним чином: перше джерело формування пар — неперекривні фрагменти тексту з одного опису роботи; друге джерело формування пар — “профіль” та “роль” розділи однієї вакансії; третє джерело формування пар — анотація та останній опис роботи з резюме.

4. Запропоновано новий метод автоматичного створення датасету вакансія-резюме, який полягає у використанні структури документа і визначеного опису останньої ролі та перетворення цього запису на опис вакансій з використанням великої мовної моделі.

5. Запропоновано метод скорочення тексту з урахуванням структури документу та ключових фраз. Цей метод полягає у скороченні кожної секції

пропорційно до її відсоткового внеску у загальну довжину резюме на основі виділення ключових фраз.

6. Розроблено метод некерованого попереднього тренування для сумаризації документів у сфері управління персоналом. Цей метод полягає у використанні секції “анотація” з резюме для некерованого тренування моделі сумаризації, а також у застосуванні зваженої функції втрат, яка підвищує вагомість для токенів, які формують ключові фрази.

3 ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПРЕДМЕТНО-ОРІЄНТОВАНОГО АНАЛІЗУ ПРИРОДНОМОВНИХ ТЕКСТІВ

3.1 Сфера застосування інформаційної технології та її методологія

На основі запропонованих в роботі методів створено інформаційну технологію предметно-орієнтованого аналізу природномовних текстів, яку назвали “AI ResJobFit”. Метою цієї технології є оптимізація процесів підбору та рекомендацій резюме до вакансій. Передбачається, що користуватися даною інформаційною технологією будуть рекрутери, відповідальні за підбір кандидатів.

Сфера застосування технології — це створення рекомендації та фільтрації кандидатів щодо вакансій на етапі відбору кандидатів. В даній технології знайдено заміну використанню комп’ютерного зору для обробки візуально насичених документів, що значно підвищує швидкість обробки документів. На відміну від класичних методів контекстно-освіченого подання ключових фраз, в розробленій технології застосовано нові методи та алгоритми, які дозволяють використовувати лише одну модель та не вимагають обчислення подань для кожної фрази шляхом “прогону” того ж фрагмента тексту з іншою маскою через модель.

Технологія підбору резюме щодо вакансій “AI ResJobFit” може бути застосована, як показано на рисунку 3.1.

Технологія аналізу резюме та зіставлення з вимогами вакансій в умовах відсутності рекрутера може бути проведена як з етапом сумаризації, так і без нього (рисунок 3.2, А). В цій технології роль людини мінімальна та зводиться до використання результатів. Це дозволяє не використовуючи працю рекрутерів надавати результати безпосередньо менеджерам для прийняття рішень.

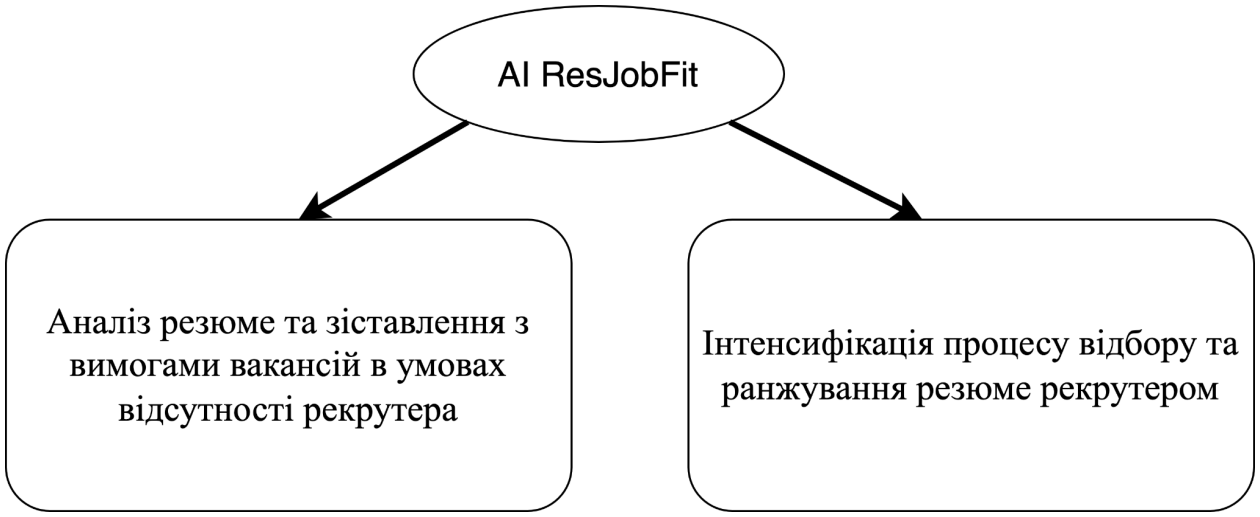


Рисунок 3.1 — Сфера застосування інформаційної технології предметно-орієнтованого аналізу природномовних текстів AI ResJobFit

Технологія інтенсифікації процесу відбору та ранжування резюме рекрутером може бути проведена як із застосуванням сумаризації так і без застосування сумаризації (рисунок 3.2, В). Ця технологія дає можливість рекрутерам швидко та зручно ознайомлюватися з рекомендованими кандидатами та відфільтровувати їх. Тобто результати роботи цієї технології придатні для подальшого опрацювання людиною. За допомогою запропонованої технології праця людини значно полегшується та швидкість обробки значно підвищена в порівнянні з традиційними технологіями.

За технологією інтенсифікації процесу відбору та ранжування резюме рекрутером, процес відбору дає рекрутерам можливість швидко та зручно ознайомлюватися з рекомендованими кандидатами, що оптимізує їх працю.

Схематично структурно-процесну модель інформаційної технології AI ResJobFit надано на рисунку 3.3.

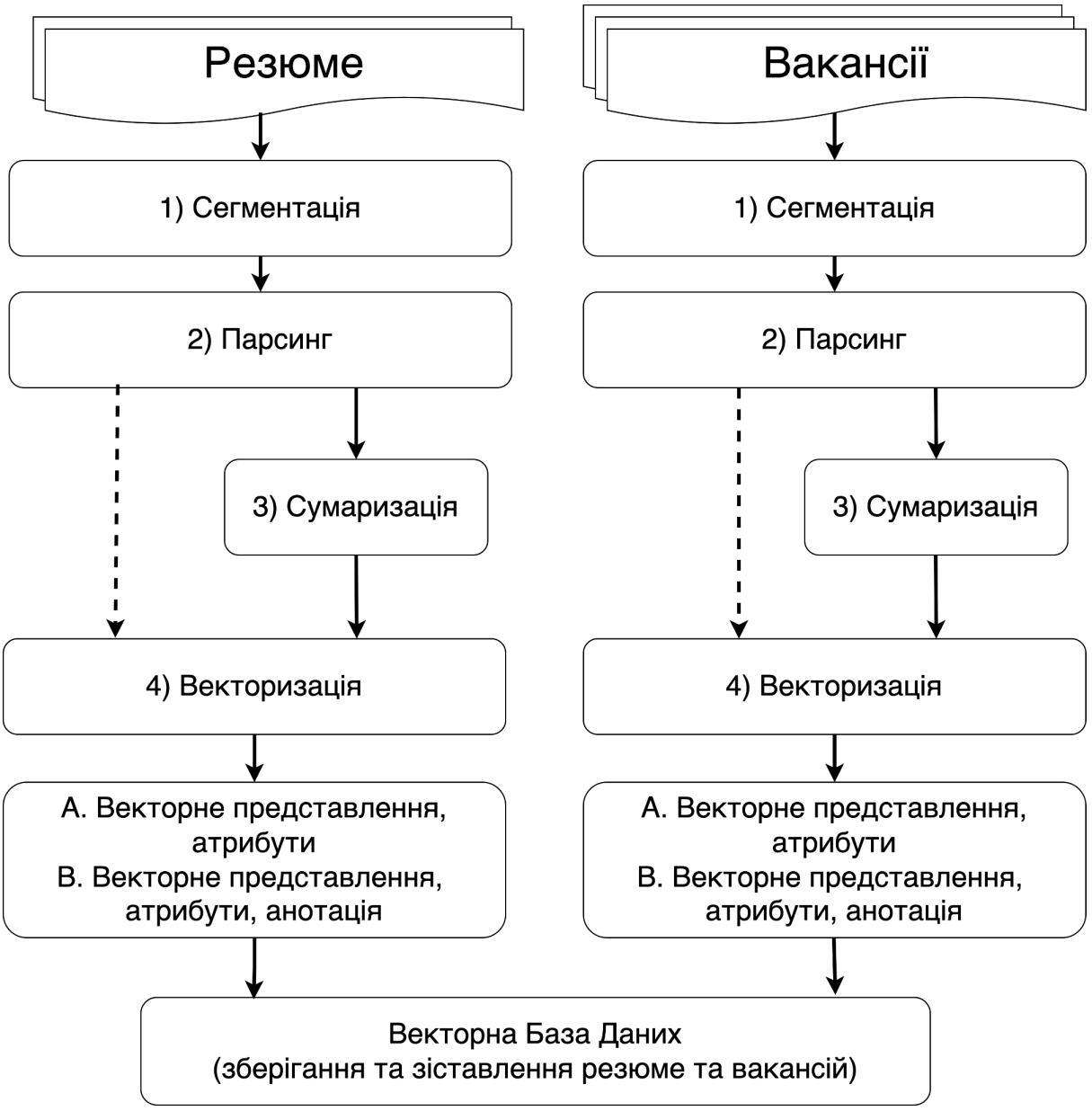


Рисунок 3.2 — Схема інформаційної технології предметно-орієнтованого аналізу природномовних текстів AI ResJobFit:

A — інформаційна технологія аналізу резюме та зіставлення з вимогами вакансій в умовах відсутності рекрутера

B — інформаційна технологія інтенсифікації процесу відбору та ранжування резюме рекрутером

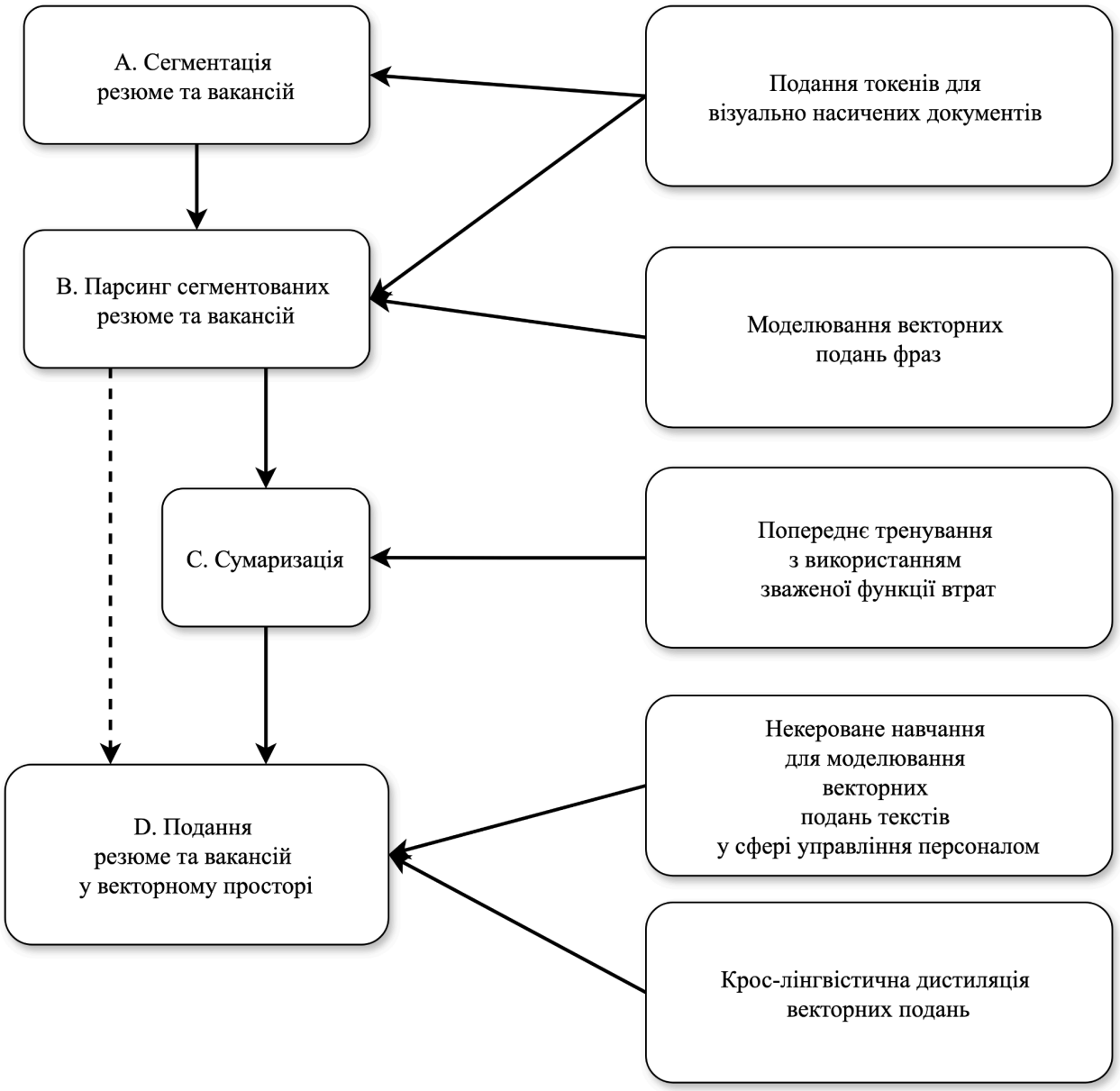


Рисунок 3.3 — Структурно-процесна модель інформаційної технології AI ResJobFit

Кожну з моделей, використану на вищезазначених етапах, піддавали прийому “дистиляція” за класичною методикою [57]. Це дозволило значно пришвидшити моделі без погіршення їх якості. Зрозуміло, що це наклало додаткові вимоги до дистиляційного тренування, а також тренування моделі-вчителя, але так як ці додаткові вимоги стосуються лише етапу тренування моделі і не вимагають їх застосування на етапі використання технології, то проведення дистиляції є доцільним.

3.2 Компоненти технології обробки природномовних текстів

Технологію предметно-орієнтованої обробки природномовних текстів у сфері управління людськими ресурсами назвемо “AI ResJobFit”. Ця технологія є послідовністю наступних етапів, описаних блоками “Сегментація”, “Парсинг”, “Векторизація”, результатами виконання яких є векторні представлення та атрибути, що визначають резюме чи вакансію, які зберігаються у векторній базі даних QDrant.

Кожний з блоків було досліджено відповідно до схеми, наданої на рисунку 3.4

А. *Сегментація* передбачає поділ тексту на розділи відповідно до заздалегідь визначеного набору категорій для подальшої обробки. Для задачі резюме цей набір категорій включає: “заголовок”, “персональні дані”, “профіль та навички”, “робота”, “проекти”, “освіта”, “сертифікації”, “тренування”, “публікації”, “рекомендації”. Вакансії сегментують у 3 категорії: “компанія та переваги”, “профіль” і “роль” відповідно.

Тексти сегментують за допомогою модифікації моделі BERT, чутливої до стилів і використання великих літер [1], що описано в розділі 2.1 та 4.1.

У результаті сегментації резюме поділяють на три основні секції, які висвітлюють різні аспекти профілю кандидата:

– Мета, анотація та навички – ця група охоплює самопредставлення кандидата, його навички та кар'єрні прагнення.

– Досвід роботи – ця група надає детальний огляд попередніх ролей, обов’язків і досягнень кандидата.

– Освіта, навчання, ліцензії та сертифікати – ця група включає академічні кваліфікації, відповідне формальне навчання та ліцензії, які підтверджують кваліфікацію, зазначену в резюме.

Вакансії поділяють на дві секції, які описують кандидата:

– майбутні функції – це інформація, що стосується майбутніх функцій працівника;

– профіль – опис кваліфікацій та вимог, зазначених у вакансії.

B. Парсинг необхідний для вилучення важливих структурованих атрибутів. Серед них – дати початку та завершення роботи, освітні ступені, сертифікати, фрази-навички, назви посад тощо. Для парсингу використовувалася архітектура CapStyleBERT [1].

Проводиться нормалізація фраз-навичок та назв посад, що детально представлено в розділах 2.2, 2.3 та 4.2, 4.3.

C. Сумаризація відіграє важливу роль у наданні рекрутерам стислого огляду кандидата, а також у виділенні найбільш релевантних деталей для процесу зіставлення. Цей етап є важливим для підвищення якості інформаційної технології “AI ResJobFit”, а також для оптимізації часу, який рекрутер витрачає на ознайомлення з резюме. Але цей етап уповільнює технологію через використання генеративної моделі, необхідної для абстрактивної сумаризації.

Процес сумаризації організовано, як показано на рисунку 3.5.

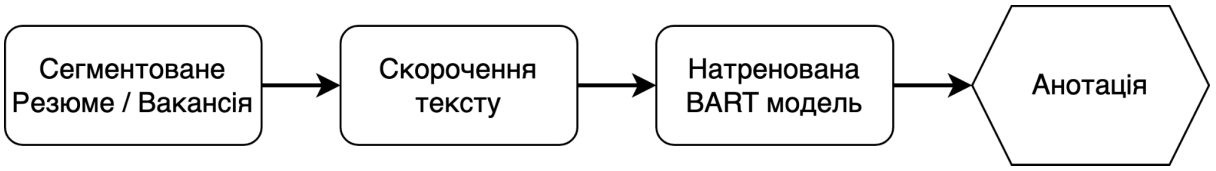


Рисунок 3.5 — Схема роботи модулю сумаризації

Через різну довжину зразків, деякі з яких перевищують максимальну межу довжини використовуваної моделі, в даній технології застосовували скорочення тексту, що детально представлено в розділах 2.4.1, та 4.4.1.

D. Векторизація потрібна для перетворення текстових даних у вектори фіксованого розміру, придатні для порівнянь. Схему цього модуля представлено на рисунку 3.6.

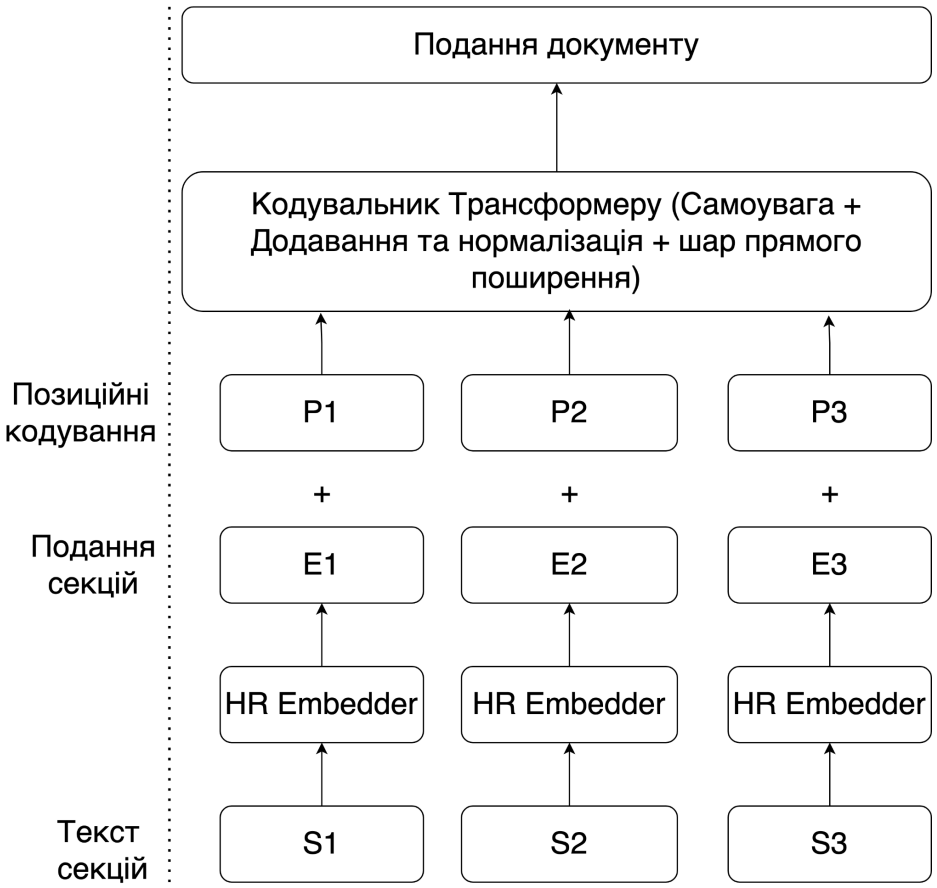


Рисунок 3.6 — Схема модулю *Векторизації*

Кожна секція визначена для надання огляду різних аспектів профілю кандидата. Після цього кожна секція кодується окремо за допомогою попередньо навченої мовної моделі, яку було визначено як HR Embedder.

Деякі резюме чи вакансії можуть не містити певних компонентів. Наприклад, деякі кандидати можуть не мати згадок про ліцензії чи сертифікації. Аналогічно, випускники або кандидати-початківці можуть не мати попереднього досвіду роботи. Такі секції будуть мати нульові вектори.

В результаті обробки секцій, модуль векторизації перетворює текст в послідовність векторних подань. Причому середнє значення подань токенів береться як подання секції. Це подання слугує унікальним “токеном” для послідовності вищого рівня, де кожен токен-подання секції робить внесок у загальне представлення профілю.

Щоб забезпечити моделі можливість створювати контекстуалізоване фінальне подання документів, в її архітектуру доцільно ввести механізм

самоуваги разом із абсолютним позиційним кодуванням. Це кодування допомагає моделі зберігати інформацію про порядок секцій, дозволяючи їй враховувати як ієрархічні, так і контекстуальні нюанси.

Математично це можна узагальнити наступними формулами:

$$h_{ij} = HR\_Embedder(t_{ij}), \quad (3.1)$$

де h_{ij} — векторне подання j-ого токену i-ої секції; t_{ij} — j-ий токен в i-ій секції.

Подання i-ої секції h_i отримуємо шляхом обчислення середнього значення векторів токенів:

$$E_i = \frac{1}{N}\sum_{j=1}^{N} h_{ij}, \quad (3.2)$$

де N — кількість токенів в i-ій секції; h_{ij} — подання j-ого токену i-ої секції; E_i — подання i-ої секції.

Послідовніть $\{E_{1,}\ E_{2,\ldots,}\ E_M\}$ збагачується абсолютним позиційним кодуванням P_i задля збереження інформації про послідовність секцій:

$$z_i = E_i + P_i\ , \quad (3.3)$$

де z_i є контекстуалізованим векторним поданням i-ї секції, що поєднує як її зміст, так і позиційну інформацію.

Механізм самоуваги обробляє послідовність $\{z_{1,}z_{2,\ldots}z_M\}$ з метою генерації контекстуалізованого представлення профілю:

$$E_{final} = Attention(\{z_{1,}z_{1,}\ldots z_M\}), \quad (3.4)$$

де E_{final} — остаточне представлення резюме чи вакансії; M це кількість секцій в документі (3 в резюме та 2 у вакансії); z_i — контекстуалізованим векторним поданням i-ї секції.

3.3 Крос-лінгвістична дистиляції векторних подань текстів

Розроблена технологія “AI ResJobFit” може бути застосована для різних мов, в тому числі, для української.

Для навчання якісних текстових подань зазвичай потрібні великі корпуси даних, які для мов із обмеженими ресурсами недоступні. Враховуючи наявність потужних моделей, здатних створювати якісні подання текстів для англійської мови, разом із парами перекладених текстів, одним із можливих рішень є дистиляція знань моделі-вчителя для англійської мови в модель-студент, яка працюватиме для мови з обмеженими ресурсами. Зокрема, крослінгвістичне вирівнювання подань дозволяє переносити знання, отримані моделлю для однієї (джерельної) мови, на іншу (цільову) мову без необхідності специфічної анотації даних для цільової мови.

Для навчання моделі подань українського тексту моделі-студенту потрібні пари текстових зразків, узгоджені з багаторесурсною (у нашому випадку — англійською). Навчальні дані зібрані з сайту Opus [80] і складаються з комбінації декількох корпусів. Зважаючи на неоднорідну якість даних та мовне забруднення, були застосовані кроки для первинної обробки текстів.

Методика підготовки даних щодо “очищення” текстів для тренування включає наступні кроки:

1) Дедуплікація — видалення повторюваних текстів

2) Ідентифікація мови та фільтрація зразків, що містять неукраїнські тексти.

3) Відсіювання вибірок, що відрізняються за довжиною більш ніж удвічі.

Для тестування запропонованої моделі переходу від англійської мови до української було розроблено еталон, який охоплює наступні домени: законопроекти, література, петиції, форум, новини.

3.4 Дистиляція моделей

Дистиляція моделей була проведена із застосуванням класичного підходу [57], адаптованого для конкретного завдання (векторні подання,

класифікація токенів тощо) із використанням традиційного підходу Кульбака-Лейблера [19].

MiniLm v2 — підхід передбачає незалежну від конкретного вирішуваного завдання дистиляцію з використанням дистиляції прихованих шарів самоуваги.

Традиційний підхід до дистистиляції для завдань класифікації чи передбачення токенів передбачає використанням розподілу вірогідностей моделі-вчителя на основі розходження Кульбака-Лейблера. Математично розходження Кульбака-Лейблера визначається наступною формулою:

$$KL(P||Q) = \sum_{x}(P(x) * log\frac{P(x)}{Q(x)}), \quad (3.5)$$

де *KL* — розходження Кульбака-Лейблера; *P(x)* — розподіл вірогідностей моделі-вчителя, *Q(x)* — розподіл вірогідностей моделі-студента.

Нульове розходження Кульбака-Лейблера означає тотожність двох розподілів вірогідностей, тоді як вищі значення означають меншу подібність двох розподілів вирогідностей.

Для дистиляції векторних подань тексту доцільно застосовувати функцію втрат на основі косинусної подібності, зіставляючи подання моделі-вчителя та моделі-студента у векторному просторі. Зважаючи на різний розмір подань для моделей різного розміру (так, модель RoBERTa large має розмірність векторних подань — 768, тоді як MiniLMv2 з архітектурою L12xH384 [57] має розмір векторних подань лише 384), модель-вчителя доцільно тренувати за допомогою “Matryoshka representation learning” підходу [80]. Цей підхід передбачає використання лише перших компонент векторного подання моделі-вчителя, що відповідають розміру подань моделі-студента, при тренуванні моделі-вчителя. Це забезпечує узгодженість векторних подань та ефективну передачу знань для подальшої дистиляції.

Адаптований метод дистиляції полягає у використанні “великої” моделі-вчителя (моделі, яка має більшу кількість слоїв та параметрів, ніж

базова версія). В якості такої моделі було застосовано RoBERTa-large [81] — кількість параметрів цієї моделі складає 355 мільйонів. В якості моделі, яка виступає в ролі цільової було використано MiniLMv2 з архітектурою L12xH384 [56] з 41 мільйонами параметрів. Обидві модели мають однаковий словник, що уможливлює дистиляцію.

Процес дистиляції має два етапи:

— Тренування моделі-вчителя;

— Тренування моделі-студента

1. *Тренування моделі-вчителя* передбачає тренування моделі відносно великого розміру на наявних якісних даних (кероване навчання)

2. *Тренування моделі-студента* передбачає тренування маленької моделі на великих обсягах даних імітувати векторні подання чи розподіл вірогідностей моделі-вчителя для кожного токену.

3.5 Методи оцінки результатів аналізу документів

3.5.1 Метрики для оцінювання класифікації

Для оцінки ефективності алгоритмів, що використовуються в задачах класифікації, широко застосовуються такі метрики, як точність (Precision), повнота (Recall) та F1-міра (F1 Score). Ці показники дозволяють проводити детальний аналіз результатів роботи моделі, враховуючи як правильність передбачень, так і їхню повноту.

Точність (Precision) обчислюють за формулою:

$$\text{Точність} = \frac{TP}{TP + FP}, \tag{3.6}$$

де TP (True Positive) — кількість елементів, які система правильно класифікувала як належні до позитивного класу; FP (False Positive) — кількість елементів, які система помилково класифікувала як належні до позитивного класу, хоча насправді вони належать до негативного класу.

Ця метрика показує, наскільки точними є передбачення системи для позитивного класу, тобто яка частка з передбачених як позитивні дійсно є правильними.

Повнота (Recall) визначає, яку частку об’єктів позитивного класу модель правильно класифікувала. Вона розраховується за формулою:

$$\text{Повнота} = \frac{TP}{TP + FN}, \tag{3.7}$$

де TP (True Positive) — кількість об’єктів, правильно класифікованих як позитивний клас; FN (помилково негативні, з англ. False Negative) — кількість об’єктів позитивного класу, які модель помилково класифікувала як негативний клас.

Висока повнота означає, що модель знаходить більшість релевантних об’єктів, але може включати також нерелевантні.

F1-міра є гармонійним середнім між точністю та повнотою, забезпечуючи збалансовану оцінку ефективності моделі. Вона розраховується за формулою:

$$F1 = \frac{2 * \text{Точність} * \text{Повнота}}{\text{Точність} + \text{Повнота}}, \tag{3.8}$$

$F1$ метрика зазвичай обирається для задач, в яких необхідно досягти балансу між точністю визначення релевантної інформації та повнотою. $F1$ є гармонійним середнім між точністю та повнотою, що дозволяє якісно оцінювати спроможності системи робити передбачення. Ця метрика зазвичай використовуються в задачах класифікації, а також маркування послідовностей. Ці тестові статистики можна обчислювати на двох рівнях деталізації: на рівні слів або на рівні токенів.

$$F1_{micro} = \frac{2 * \sum\limits_{\text{всі класи}} TP}{2 * \sum\limits_{\text{всі класи}} TP + \sum\limits_{\text{всі класи}} FP + \sum\limits_{\text{всі класи}} FN}, \tag{3.9}$$

де $\sum_{\text{всі класи}} TP$ — загальна кількість правильно класифікованих об’єктів, $\sum_{\text{всі класи}} FP$ — загальна кількість помилково позитивних (екземпляри, які були помилково класифіковані як такі, що належать до класу), $\sum_{\text{всі класи}} FN$ — загальна кількість помилково негативно передбачених об’єктів (екземпляри класу, які були помилково класифіковані як такі, що не належать до нього).

3.5.2 Метрики для оцінювання ранжування

Recall@N — метрика, яка часто використовується в задачах інформаційного пошуку та рекомендаційних систем для оцінки ефективності ранжування. Вона показує, яка частка релевантних елементів з усієї релевантної множини знаходиться серед топ-N результатів. Ця метрика особливо актуальна, коли кількість релевантних документів маленька.

Середня точність (MAP — з англ. mean Average Precision) — метрика вимірює середню частку релевантних документів з топ-N результатів для кожного з запитів. Кількість документів топ-N визначається відповідно до конкретного завдання, але часто може бути визначена як кількість релевантних документів для кожного запиту.

Математично середня точність обчислюється за наступною формулою:

$$MAP = \frac{\sum_{1}^{Q} AP}{Q}, \tag{3.10}$$

де Q — кількість запитів у корпусі, AP — середня точність для одного з запитів, яка розраховується за формулою:

$$AP = \frac{TP}{K}, \tag{3.11}$$

де K — кількість релевантних документів, які належать до запиту, TP — кількість правильно передбачених документів

Середньо зворотній ранг (MRR — з англ. mean Reciprocal Rank) — метрика, яка вимірює середній показник зворотніх рангів першого коректно передбаченого документа для кожного запиту

$$MRR = \sum_{q=1}^{Q} \frac{\frac{1}{rank_{qi}}}{Q}, \quad (3.12)$$

де $rank_{qi}$ — позиція першого правильно передбаченого документа в ранжованому списку документів для запиту q. Q — кількість запитів у корпусі.

Нормалізований дисконтований накопичений прибуток (NDCG — з англ. normalized Discounted Cumulative Gain) — вимірює якість ранжування. Обчислюється як частка між дисконтований накопиченим прибутком (DCG) та DCG у випадку ідеального ранжування.

$$DCG@K = \sum_{k=1}^{K} \frac{rel_i}{log_2(i+1)}, \quad (3.13)$$

де rel_i — релевантність документу щодо запиту — може набувати значення 1 чи 0 в залежності від того, чи правильне передбачення, K — зазвичай встановлюється рівних кількості релевантних документів для запиту.

$$nDGC = \frac{DCG@K}{IDGC@K}, \quad (3.14)$$

3.5.3 Метрики для оцінювання згенерованого тексту

Найбільш розповсюдженою метрикою для оцінювання якості сумаризацій виступає ROUGE (з англ. Recall-Oriented Understudy for Gisting Evaluation). Ця метрика вимірює відношення кількості N-грамів, які перекриваються між згенерованим та референтним текстами.

$$ROUGE_N = \frac{\sum Count_{match}(Ngram)}{\sum Count(Ngram)}, \quad (3.15)$$

де N — довжина N-граму (часто встановлюється в розмірі 1-3); $Count_{match}(Ngram)$ — кількість N-грамів які перекриваються в згенерованому та референтному тексті; $Count(Ngram)$ — кількість N-грамів в референтному ($ROUGE_{recall}$) та згенерованому ($ROUGE_{precision}$) тексті.

$$ROUGE_{F1} = \frac{2 * ROUGE_{precision} * ROUGE_{recall}}{ROUGE_{precision} + ROUGE_{recall}}, \quad (3.16)$$

ROUGE-L — метрика, яка вимірює співвідношення довжини найдовшої спільної підпослідовність (LCS — з англ. Longest Common Subsequence) по відношенню до довжини згенерованого текста.

$$ROUGE_L = \frac{LCS}{len(summary)}, \quad (3.17)$$

де LCS — довжина найдовшої спільної підпослідовність між згенерованим та референтним текстом; $len(summary)$ — довжина згенерованого тексту.

Висновки за розділом 3

1. Інформаційна технологія предметно-орієнтованого аналізу природномовних текстів AI ResJobFit може бути застосована за двома напрямками: вироблення рекомендацій резюме в умовах відсутності рекрутера, та інтенсифікації процесу відбору та ранжування резюме рекрутером, що дає можливість рекрутерам швидко та зручно ознайомлюватися з рекомендованими кандидатами та відфільтровувати їх.

2. Технологія “AI ResJobFit” є послідовністю застосування наступних етапів: “Сегментація”, “Парсинг”, “Сумаризація”, “Векторизація”. В результаті застосування цих етапів документ перетворюється на сукупність

атрибутів, анотації та векторного подання, які зберігаються у векторній базі даних QDrant для подальшого зіставлення резюме та вакансій.

3. Встановлено, що для оцінювання етапів технології “AI ResJobFit” необхідно обчислювати наступні показники: *F1, Recall@N, MAP, MRR, nDCG,* $Rouge_N$

4 РЕАЛІЗАЦІЯ МЕТОДІВ ОБРОБКИ ПРИРОДНОМОВНИХ ДОКУМЕНТІВ ДЛЯ АНАЛІЗУ ТЕКСТІВ В СФЕРІ УПРАВЛІННЯ ПЕРСОНАЛОМ

4.1 Подання токенів для візуально насичених документів

Дані, використані в цьому дослідженні, отримані з бази резюме Daxtra Technologies і включають резюме в різних форматах (pdf, doc, docx, rtf).

Інформація про текст і стиль з файлів резюме було витягнуто за допомогою пропрієтарного фреймворку, який дозволяє правильно ідентифікувати розміри шрифтів, стилі, кольори та геометричне розташування кожного слова.

Нормалізації символів нового рядка було проведено так, щоб досягти умови не більше двох послідовних символів нового рядка. Всі розділові знаки та стоп-слова були залишені, оскільки попередньо навчені моделі, такі як BERT і GPT, демонструють здатність ефективно розуміти та обробляти текст із цими елементами, що дозволяє проводити більш повний аналіз природномовних текстів.

4.1.1 Експериментальні дані

Було використано приблизно 2 мільйони резюме для попереднього навчання. Щоб зменшити витрати на обчислення, ваги моделі були ініційовані з попередньо навченої версії BERT (base-uncased). Після цього було проведено некероване попереднє навчання з використанням MLM (Masked Language Modelling) стратегії навчання.

Добре відомо, що доменне адаптивне попереднє навчання покращує якість подань токенів. Тому для справедливого порівняння з запропонованим підходом було проведено попереднє некероване навчання з оригінальної

BERT моделі без використання інформації про стиль і капіталізацію на тих самих даних, що і для запропонованої архітектури.

Для оцінювання моделей було відібрано резюме, які належать до різних професійних секторів і мають різні формати файлів. Розподіл форматів отриманого набору даних (датасету) (7 219 резюме) показано на рисунку 4.1.

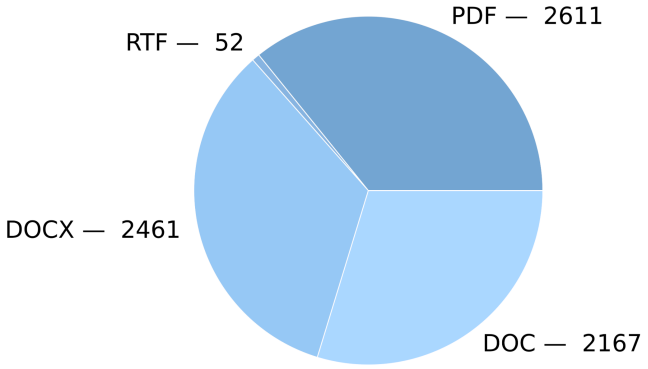


Рисунок 4.1 — Розподіл форматів у датасеті

Для ідентифікації сегментів текстові фрагменти резюме були розподілені у такі 10 класів: “заголовок”, “персональні дані”, “профіль та навички”, “робота”, “проекти”, “освіта”, “сертифікації”, “тренування”, “публікації”, “рекомендації”. Усі токени, які не належали до вказаних класів, залишалися без анотацій та отримували тег “O”.

На рисунку 4.2 показано розподіл токенів, призначених кожному з класів.

Завдання сегментації було розглянуто як задачу маркування послідовностей, подібну до підходу NER (визначення іменованих сутностей), і застосовано схему тегування BIO, де “B-” позначає початок відповідного сегмента, “I-” позначає внутрішню позицію токена, а “O” використовується для позначення токенів, що не належать до жодного значущого класу.

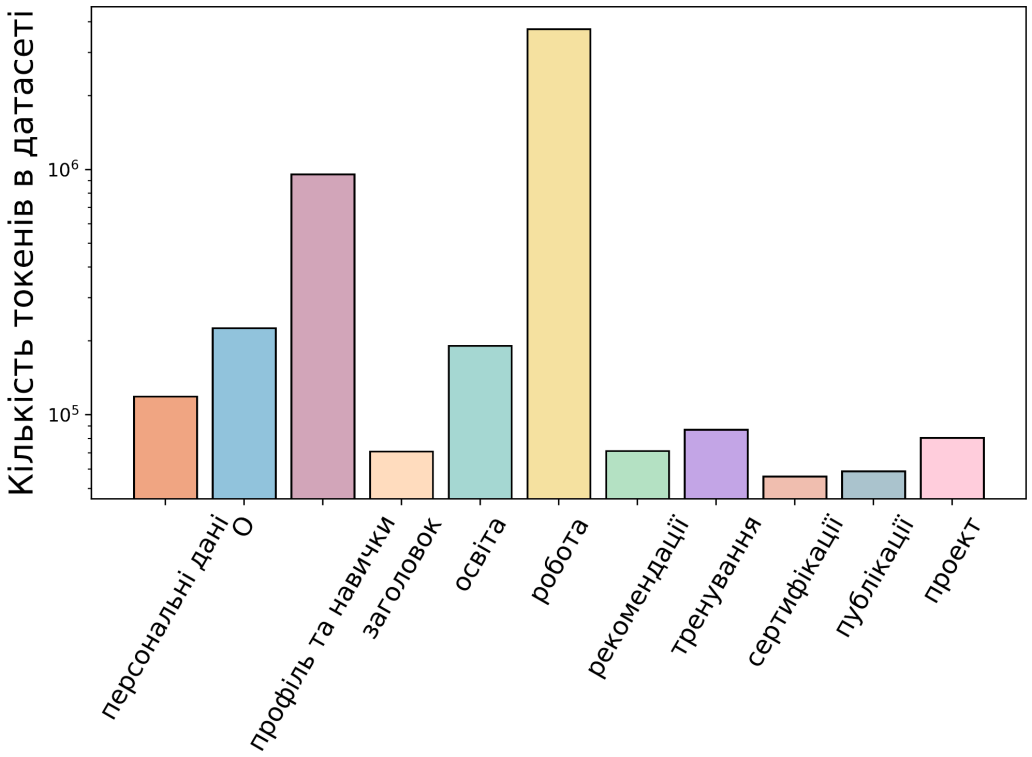


Рисунок 4.2 — Розподіл класів у датасеті

Описана вище схема тегування BIO створює дисбаланс у датастеті. Це відбувається через те що, на відміну від задачі тегування іменованих сутностей, які зазвичай короткі, сегменти резюме часто включають в себе декілька рядків чи навіть абзаців. Тому всі наступні токени в тому ж рядку, починаючи з токена “B-”, також отримують мітку “B-” із відповідним тегом. Це дозволяє знизити дисбаланс у датасеті.

Датасет було поділено на тренувальну, валідаційну та тестову вибірки на основі документів у таких пропорціях: 70 : 10 : 20 відповідно, що призвело до 5 052 документів у тренувальному наборі, 722 у валідаційному та 1 445 у тестовому.

4.1.2 Практичне застосування і оцінка результатів

Усі експерименти проводилися на 1 машині з GPU-карткою A100. У нашій конфігурації навчання ми налаштували швидкість навчання на рівні

$2 \cdot 10^{-5}$ і застосували стратегію лінійного розігріву протягом початкових 10% кроків навчання. Крім того, був використаний оптимізатор AdamW для підвищення ефективності навчання. Формат з половинною точністю чисел з плаваючою комою (FP16) використовувався для прискорення навчання та економії обчислювальних ресурсів.

Враховуючи розподіл текстових фрагментів у 10 класів, для покращення читабельності та зручності аналізу результатів ми наводимо F1-мікро метрику.

Щоб дослідити, як моделі працюють в умовах обмежених ресурсів, було оцінено кожну модель, використовуючи навчальні дані в діапазоні від 5% до 100% доступного навчального набору.

Як базові було використано наступні моделі:

— bert-base-cased і bert-base-uncased — оригінальні моделі, представлені в статті [22];

— bert-base-uncased_pt — модель, яка було додатково адаптована до домену управління персоналом шляхом некерованого навчання з MLM тренувальною стратегією

На рисунку 4.3 візуалізовано метрики, досягнуті кожною з моделей.

Модель BERT у версії з урахуванням регістру (cased) демонструє найгіршу продуктивність, так як у багатьох випадках капіталізація використовується для різних цілей, зокрема для виділення власних назв, таких як навички та навчальні заклади. З цього випливає, що покладатися виключно на капіталізацію як на маркер меж розділів може бути недостатньо надійним та послідовним підходом. Крім того, модель BERT-cased розглядає слова з різною капіталізацією як окремі токени, що збільшує розмір словника та може призводити до проблем із розрідженістю даних під час навчання. Це ускладнює процес навчання моделі та її здатність до ефективного узагальнення.

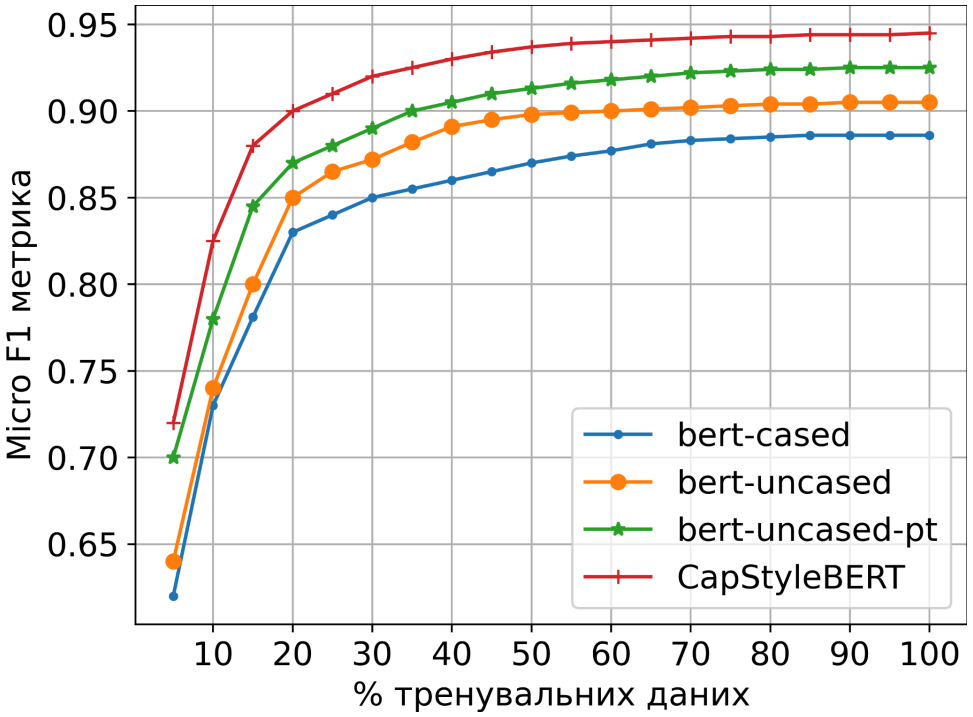


Рисунок 4.3 — Точність моделей при різних обсягах даних в навчальному наборі

Як видно з рисунку 4.3, застосування некерованої адаптації до домену з використанням стратегії MLM для базової моделі виявляється особливо корисним у сценаріях з обмеженими ресурсами. Ця техніка призводить до значного покращення продуктивності на 5,4% у випадках, коли розмір датасету обмежений лише 5% від доступних тренувальних даних, що еквівалентно датасету з низькою кількістю резюме (5% від тренувального датасету складає 253 резюме для даного дослідження).

Із поступовим збільшенням розміру тренувальних даних значний приріст продуктивності, досягнутий завдяки цьому підходу, поступово зменшується. Фактично, коли використовується весь обсяг тренувальних даних, що становить 5 052 резюме, спостережуване покращення продуктивності скорочується до 0,2%.

Таким чином, некерована адаптації до домену в поєднанні зі стратегією MLM є особливо ефективною при обмежених ресурсах даних.

Розроблена CapStyleBERT модель демонструє кращу продуктивність при всіх розмірах датасетів, перевершуючи моделі, які не використовують інформацію про форматування тексту, на 3-6% за різної кількості тренувальних зразків.

Подібним чином було досліджено можливість сегментації вакансій у три класи: “вимоги”, “роль” та “опис компанії та переваг”, а також можливість виділення ключової інформації з різних розділів резюме та вакансій з використанням представленої архітектури.

При сегментації вакансій використовувалося 1 500 анотованих людиною вакансій, середній показник якості (F1 мікро метрика) склала 0,93 і перевершила показники якості адаптованої до домену версії моделі BERT, навченої за допомогою MLM стратегії тренування, на 1%. Отримані результати якості витягнення ключової інформації становлять 0,79-0,98 F1 метрика на відповідних датасетах, створених для кожної з секцій. Ці датасети складалися з розмічених людиною даних і включали 1-5 тисяч анатованих секцій. Застосування розробленої CapStyleBERT моделі перевершує показники адаптованої до домену версії моделі BERT, навченої за допомогою MLM стратегії тренування, на 2-9%.

4.2 Модель векторного представлення фраз у контексті для їх групування та нормалізації

4.2.1 Експериментальні дані

Відповідно до [66], для створення датасету еталонного було зібрано оголошення про роботу з великого державного веб-сайту з працевлаштування myfuturejobs.gov.ua [83], який містить не лише назви вакансій та відповідні описи, але й нормалізовану форму назви вакансії, що відповідає вимогам ESCO [84]. Цю категорії обирає автор оголошення, що дає можливість використовувати її як еталон. З датасету було видалено оголошення про

вакансії, що містять неанглійські описи. Статистика остаточного набору даних, використаного для тестування, показана на рисунку 4.4. Цей набір даних було оприлюднено [2].

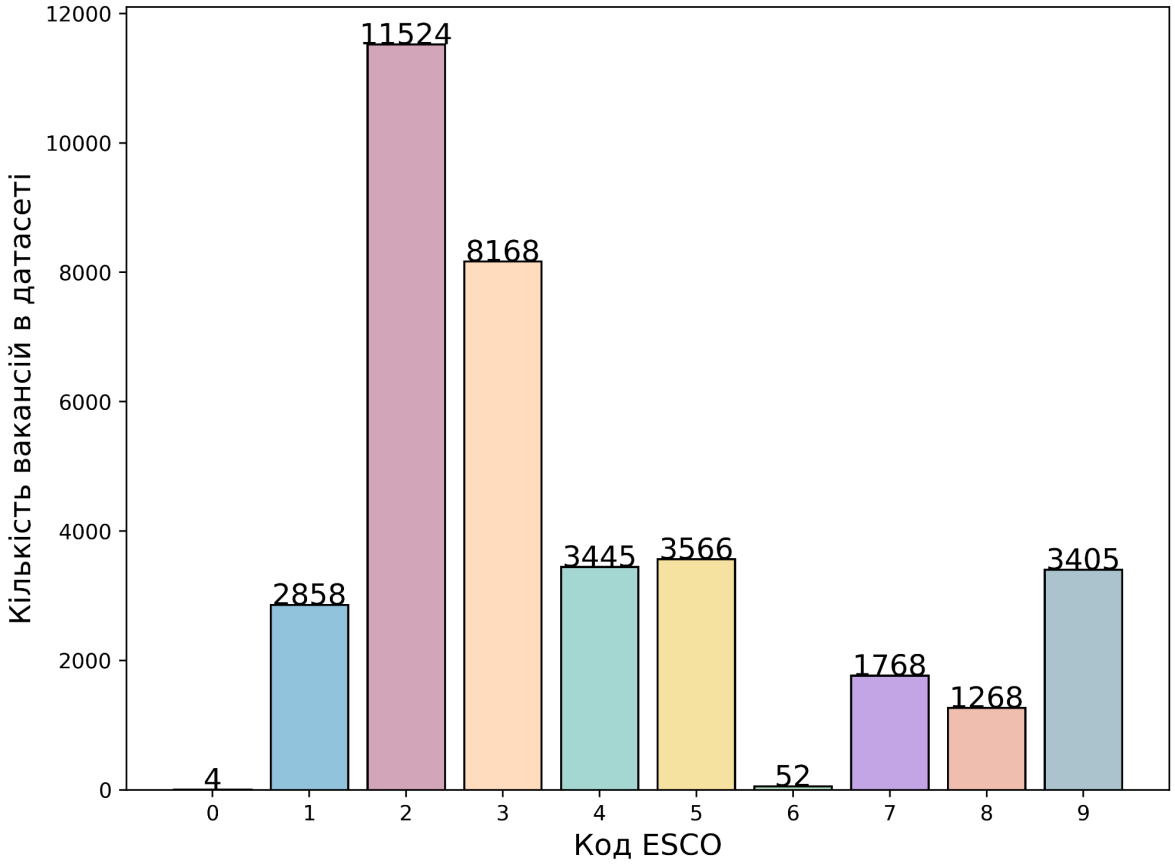


Рисунок 4.4 — Кількість оголошень про роботу в зібраному датасеті

** де ESCO код відповідає:*

0 — Професії збройних сил

1 — Менеджери

2 — Професіонали

3 — Фахівці та допоміжні фахівці

4 — Офісні допоміжні працівники

5 — Працівники сфери послуг та торгівлі

6 — Кваліфіковані робітники сільського, лісового господарств та рибальства

7 — Кваліфіковані працівники ремісничих і подібних професій

8 — Оператори та складальники устаткування і машин

9 — Найпростіші професії

Як видно з рисунка 4.4, зібраний датасет не є збалансованим: вакансії переважно розміщені для сімейств професій “Професіонали” та “Фахівці та допоміжні фахівці”, тоді як “Кваліфіковані робітники сільського, лісового господарств та рибальства”, а також “Професії збройних сил” майже відсутні.

При роботі з незбалансованими наборами даних вирішення завдання як задачі класифікації може бути недоцільним, оскільки клас меншості може бути легко затьмарений класом більшості [76]. З іншого боку, завдання ранжування може бути кращим варіантом для таких наборів даних. Ранжування — це метод, який фокусується на відносному впорядкуванні вибірок, а не на абсолютній класифікації. Його мета — відрізнити реальну позитивну пару від усіх інших можливих комбінацій з іншими зразками. Цей метод ефективний для незбалансованих наборів даних, оскільки не вимагає рівної кількості позитивних і негативних прикладів, що робить його більш гнучким і відповідним реальним сценаріям, де дані зазвичай незбалансовані. Крім того, цілі ранжування можуть також надавати пріоритет правильному ранжуванню неправильно класифікованих прикладів, тим самим покращуючи загальну продуктивність моделі [85].

Для створення тренувального набору даних було використано великий навчальний корпус з ~10 мільйонів англомовних вакансій, зібраних з різних американських та британських платформ для пошуку роботи, а також ~40 мільйонів анонімізованих секцій “трудова діяльність” резюме з внутрішнього “озеру даних” компанії Daxtra Technologies.

Характеристика зібраного датасету показана в таблиці 4.1.

Як можна помітити, як резюме, так і вакансії містять дуже різноманітні назви посад. Резюме, як правило, мають набагато більш уніфіковані назви посад порівняно з тими, що зустрічаються у вакансіях, оскільки в оголошеннях часто є багато додаткової інформації, такої як переваги роботи, локації тощо.

Таблиця 4.1 — Характеристика тренувального набору даних

Джерело даних	Кількість унікальних текстів	Кількість унікальних назв посад	Середня кількість текстових описів на одну назву посади
Резюме	40M	6,5M	6,17
Вакансії	10M	4,8M	2,08

За допомогою регулярних виразів було проведено видалення URL-адресів, електронних пошт та номерів телефонів. Задля очищення зібраних вакансій від нерелевантної інформації було застосовано сегментацію, як описано в розділі 2.1 та 4.1. Це дозволило витягувати навички лише з релевантних розділів (і запобігло забрудненню навичок, які могли б бути витягнути з розділів, що містять описи рекламованих переваг та іншої нерелевантної для нашого завдання інформації).

Для ідентифікації навичок використовували модель, яка вирішує це завдання як завдання класифікації токенів. Архітектура моделі описана в розділах 2.1 та 4.1.

4.2.2 Практичне застосування і оцінка результатів

Аналіз витягнутих навичок показує, що більшість вакансій містить від 10 до 100 навичок, при цьому лише 8,6% містять менше ніж 10 і лише близько 12,4% містять більше ніж 100 навичок.

Кількість навичок на вакансію у відсотках показана на рисунку 4.5.

Максимальна довжина об'єднаних навичок встановлена на рівні 128 токенів, оскільки було виявлено, що для більшості вакансій та описів роботи цієї довжини достатньо, щоб успішно включити всі наявні навички.

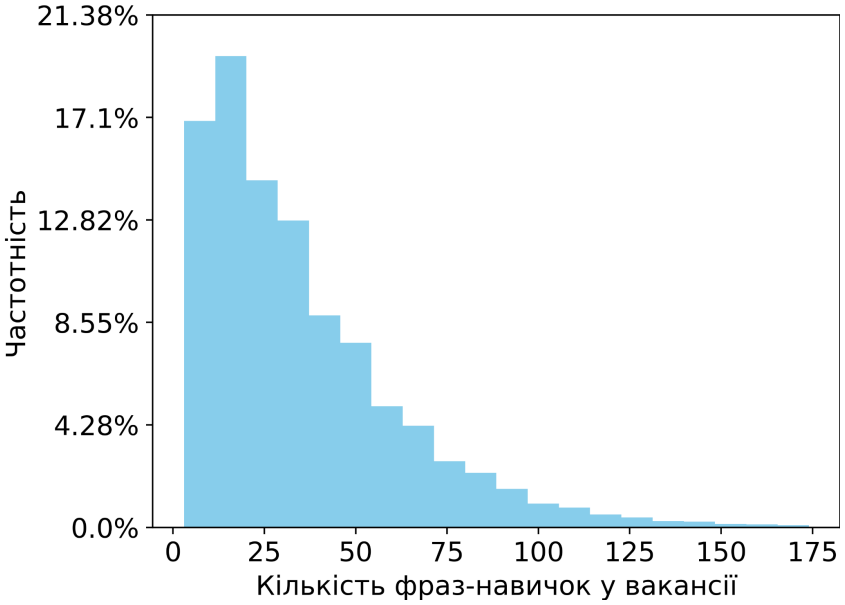


Рисунок 4.5 — Кількість фраз-навичок в одному описі роботи

Для тренування використовували функцію множинних негативних втрат (Multiple Negative Ranking Loss) [31] з негативами в батчі, а розмір батчу був встановлений у розмірі 32. Модель тренували протягом однієї епохи, при цьому проводили моніторинг ефективності на валідаційному наборі.

Результати експерименту наведено в Таблиці 4.2.

Таблиця 4.2 — Нормалізація назв посад

Модель	Recall@1	Recall@5	Recall@10
FastText	0,134	0,223	0,279
BERT*	0,142	0,180	0,254
SBERT	0,219	0,320	0,444
SBERT (з навичками)	0,233	0,361	0,50
JobBERT**	0,225	0,386	0,46
VacancySBERT (без навичок)	0,271	0,402	0,489
VacancySBERT (з навичками)	0,301	0,425	0,556

Примітка:

** Подання назви посади обчислюється шляхом усереднення подань усіх токенів назви посади.*

*** Метрики базуються на значеннях, наведених у [66], оскільки модель не була опублікована у відкритий доступ.*

Для порівняння запропонованої моделі з найкращим конкурентом (JobBERT) і визначення покращення (Δ) для двох налаштувань (з використанням або без використання навичок як додаткових ознак) використовується така формула:

$$\Delta = \frac{\sum_{N=1,5,10} \frac{R@N_{V.B.} - R@N_{J.B.}}{R@N_{J.B.}}}{3}, \quad (4.1)$$

де *R@N* відповідає *Recall@1*, *Recall@5*, *Recall@10* у Таблиці 4.2, *V.B.* — відповідає VacancySBERT (з навичками або без них) у Таблиці 4.2, а *J.B.* — JobBERT відповідно.

Розрахунки показують покращення на 10% і 21,5% при використанні VacancySBERT та VacancySBERT (з урахуванням навичок) відповідно.

Таким чином, зіставлення назви посади з навичками з опису роботи при тренуванні моделі дозволяє суттєво покращити результати порівняно з базовими моделями (FastText, BERT, SBERT, JobBERT). Використання навичок як додаткових ознак під час інференсу посилює отриманий ефект. Можна відзначити, що додавання навичок навіть до попередньо натренованої SBERT покращує прогнозувальну здатність моделі.

4.3 Моделювання фраз-навичок у контексті

Нова архітектура Phrase2Vec щодо вивчення подань фраз для текстів, пов'язаних з управлінням персоналом, представлена у розділі 2.3, базується на BERT з додаванням спеціальних токенів, які маркують межі фраз. Цю архітектуру було протестовано у двох сценаріях. Перший – повністю без

нагляду, де було використано приклади однієї і тієї ж фрази, що з'являються в різних контекстах, як позитивні приклади для контрастивного навчання. І другий, з напівкерованим сценарієм навчання, де для створення позитивних прикладів було використано сигнал від попередньо навченої великої мовної моделі для моделювання лексичної подібності подань речень.

4.3.1 Експериментальні дані

Для підготовки навчальних даних для експериментів, через неоднорідну якість проаналізованих резюме та вакансій, необхідно налагодити процес ретельної фільтрації. Дотримуючись класичної послідовності дій для попередньої підготовки природномовних текстів NLP, було реалізовано наступну серію етапів для очищення набору даних:

— ідентифікація та фільтрація мови;

— сегментація даних;

— дедуплікація даних;

— фільтрація зразків;

— виділення ключових фраз-навичок.

Ідентифікація та фільтрація мови зосереджена на видаленні неанглійських текстів. Оскільки дослідження було зосереджено на англійській мові, було використано попередньо навчений детектор мови fasttext [30] для фільтрації неанглійських текстів. Відфільтрованими були всі резюме та вакансії, де впевненість моделі в тому, що текст є англійським, була нижче за 80%. 6% вакансій та 3% резюме були відфільтровані на цьому етапі.

Сегментація даних полягає у виділенні абзаців, які потенційно містять ключові фразі-навички. При обробці вакансій було видалено абзаци, які не пов'язані з вимогами до кандидатів (розділи про компанію та опис переваг в оголошеннях про роботу). Сегментація була проведена з використанням моделі, яка описана в розділі 4.1. В результаті було отримано 2,8 млн речень.

Резюме були сегментовані і вилучені лише абзаци, які належали до робочих історій та розділу “анотація” (summary). В результаті було отримано 3,4 млн абзаців.

Дедуплікація даних покликана відфільтрувати повторювані зразки. Як резюме, так і дані про вакансії містять багато ідентичних або майже ідентичних текстів (через те, що люди подають одне й те саме резюме кілька разів, а рекрутери репостять одне й те саме оголошення про вакансію на кількох сайтах, а також часто використовують загальні описи чи вимоги). Поява дублікатів у навчальних даних може значно погіршити продуктивність і точність мовної моделі. Зокрема, навчання мовних моделей на дедуплікованих даних призводить до зменшення кількості кроків навчання, необхідних для досягнення такого ж або вищого ступеня точності [85].

Задля подолання дублікації було видалено всі повторювані зразки з набору даних. Задля цього було використано методику хеш-значень [86]. Ця методика виявилася доцільною через значний обсяг набору даних. Перед процедурою створення хеш-значень пробіли були нормалізовані та текст переведений в нижній регістр. Таким чином було відфільтровано 53% даних.

Фільтрація зразків необхідна для забезпечення різноманітності даних. Для цього було використано внутрішню таксономію професійних секторів Daxtra Technologies, яка складається з 28 верхніх рівнів. Зібрані речення та абзаци були автоматично класифіковані згідно з цією таксономією, після чого було проведено балансування кількості зразків. Це було зроблено шляхом видалення частини зразків, які належали до найбільш широко представлених професійних секторів (зокрема, IT).

Виділення ключових фраз-навичок з речень та абзаців з використанням архітектури моделі, було проведено як описано в розділі 4.1. Ця модель працює на рівні токенів. Речення та абзаци, які не містили хоча б однієї фрази-навички були відфільтровані.

В результаті було отримано ~2,1 млн валідних для поставленої задачі зразків. Рисунок 4.6 графічно ілюструє розподіл професійних секторів у

кінцевому навчальному наборі.

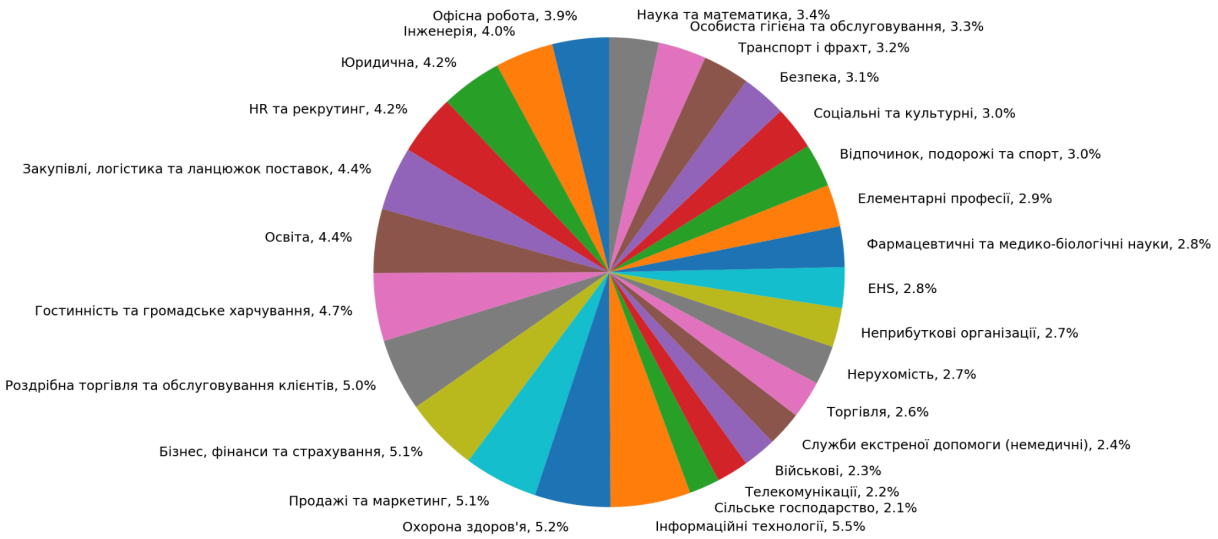


Рисунок 4.6 — Розподіл професійних секторів тренувальних даних

Для тренування моделі були відібрані навички, які з’явилися в текстах не менше 10 разів (для того щоб забезпечити наявність тренувальних пар та видалити випадкові або занадто специфічні навички)

Задля тестування підходів необхідно мати еталонний датасет, який би відображав реальну варіативність і складність навичок, що зустрічаються у текстах вакансій та резюме, і дозволяв адекватно оцінювати точність та ефективність подання цих навичок. Створення розміченого вручну набору даних достатнього розміру для нормалізації навичок неможливе при використанні дрібнозернистих таксономій (в ESCO понад 13 тисяч навичок). Використання більш грубих таксономій, таких як ONET, де є лише 35 груп навичок, не є корисним для реального застосування. Але можна створити еталонний набір, який, подібно до добре відомих наборів тестових даних для загального навчання подібності речень (STS тощо), дозволить порівнювати фрази, які описують навички, між собою.

З підготованого набору було випадковим чином відібрано 50 тисяч

фраз-навичок, які зустрічалися хоча б 10 разів, а потім згруповано їх навколо найбільш частотних 10 тисяч з цієї вибірки. Групування було проведено з допомогою евристик та декількох моделей, кожна з яких фокусується на різних аспектах, таких як лексична, морфологічна або контекстуальна схожість.

Для вибору пар для анотування були враховані певні евристики, включаючи передбачений професійний сектор, щоб визначити потенційні неправильні кандидати для кожної з відібраних фраз запиту.

Набір даних було сформовано з урахуванням наступних особливостей:

— адверсійний лексичний збіг;

— тлумачення скорочень та багатозначних слів і фраз;

— інші складні випадки.

Адверсійний лексичний збіг: було включено фрази, які мають однакові слова, але не пов'язані між собою (“flower planting” — “plant flow”; “microsoft outlook” — “strategic outlook”). Очікується, що моделям моделі, які використовують лексичну або морфологічну подібність, буде важко розрізнити такі випадками.

Тлумачення скорочень та багатозначних слів і фраз: певні слова / фрази можуть мати кілька значень залежно від контексту. Наприклад, абревіатура “ER” може означати як “employee relations” (відносини з працівниками) або “emergency room” (відділення невідкладної допомоги), залежно від контексту.

Складні випадки: вибиралися фрази, де передбачування окремих моделей не узгоджуються між собою, вони були ретельно перевірені та анотовані, щоб зробити еталонний набір даних більш складним.

Отримані результати з впровадження еталону включають набір з 5 513 унікальних фраз, пов'язаних між собою, які утворюють 6 634 негативні та 6 723 позитивні пари, що в результаті дає 13 357 анотованих пар.

Характеристика еталонного датасету наведена в таблиці 4.3

Таблиця 4.3 — Характеристика еталонного датасету

Показник	Значення
Середня кількість слів зразка	2,8
Кількість унікальних слів	2329
Відсоток слів, які збігаються в позитивних парах	9,6%
Відсоток слів, які збігаються в негативних парах	6%
Середня кількість слів у контекстному реченні або абзаці	23

Таким чином, еталонний датасет являє собою збалансований набір з невеликою кількістю лексичного збігу в зразках (<10%).

4.3.2 Практичне застосування і оцінка результатів

Було проведено тренування BERT (базова версія) моделі з використанням графічного процесора NVIDIA T4 за 1 епоху. Розмір батчу становив 24 зразка. Також тренування з використанням графічного процесора NVIDIA A100 було проведено з розміром батчу 128 зразків. Порівняння результатів для батчів різного розміру дозволяються встановити залежність якості від розміру батчу.

Початкові 10% тренувальних кроків використовували для “прогріву” в обох налаштуваннях, після чого швидкість навчання збільшувалася лінійно до максимального значення. Був використаний оптимізатор AdamW із коефіцієнтом швидкості навчання $2 \cdot 10^{-5}$ як для T4 GPU, так і для A100 GPU.

Результати експериментів підтверджують залежність якості вивчених подань від кількості негативних прикладів у тренувальному батчі (Рисунок 4.7), що є у кореляції з літературними даними [48].

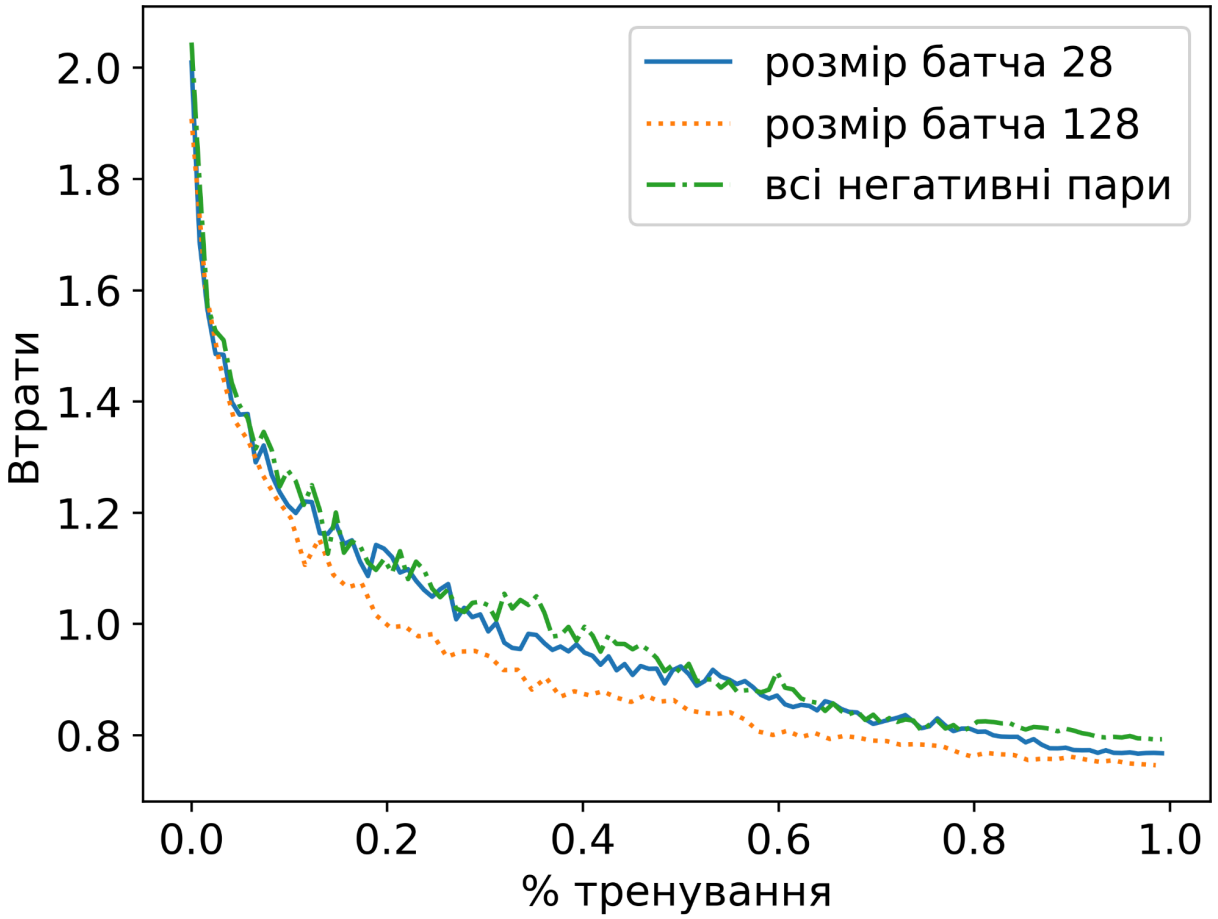


Рисунок 4.7 — Метрики втрат на валідаційному наборі даних при тренуванні

Результати досліджень щодо підхода “всі негативні пари” (Рисунок 4.7) вказують на зниження метрик на валідаційному наборі. Тому підхід “всі негативні пари” не доцільно рекомендувати для використання при тренуванні подань в умовах асиметричного датасету.

Отримані дані вказують на доцільність використання великих батчей даних (128 пар) при контрастному тренуванні. Це значення розміру батчу (128 пар) доцільно використовувати для практичних цілей.

Оскільки еталонний набір даних структуровано для вирішення проблеми парної класифікації, подання фраз необхідно розраховувати незалежно, а потім обчислювати косинусну подібність між ними. Визначення оптимального порогу подібності для кожної моделі дозволяє розрахувати середню точність, яка слугує основною метрикою в оцінці. В таблиці 4.4 наведені результати, отримані за допомогою сучасних моделей подань тексту,

які навчали без нагляду, а також результати, досягнуті за допомогою застосування представленої архітектуру Phrase2Vec.

Таблиця 4.4 — Точність, досягнута моделями на еталонному датасеті

Модель	Точність моделі натренованої на загальних даних	Точність моделі натренованої на HR даних
FastText	0,655	0,661
SimCSE	0,694	0,697
McPhraSy	не застосовується	0,706
$Phrase2Vec_{unsup}$	не застосовується	0,812

Оцінка якості векторних подань моделей, навчених без нагляду, на еталонному датасеті вказує на ефективність запропонованої архітектури. Зокрема, результати підкреслюють важливість включення контекстної інформації в процес моделювання подань навичок.

Таблиця 4.5 — Точність, досягнута на еталонному датасеті моделями, натренованими при напівкерованому сценарії навчання

Модель	Phrase-BERT	E5 base	GTE base	SimLM	Jina v1	GPT sentences aug	$Phrase2Vec_{sup}$
Точність	0,642	0,696	0,726	0,753	0,796	0,832	0,931

Таким чином, тренуванн з використанням є сигналу від попередньо натренованої моделі подань тексту є доцільним і дозволяє досягти покращення в порівнянні з повністю некерованою стратегією навчання.

Запропоноване рішення щодо моделювання контекстно-освічених подань фраз при застосуванні спеціальних токенів, які маркують текстові межі, значно перевершує показники якості в порівнянні з іншими моделями, оскільки таке рішення дозволяє моделі ефективно розпізнавати та використовувати контекстні підказки, зосереджуючись на самій фразі.

З експериментальної таблиці (Таблиця 4.5) видно, що якість репрезентацій PhraseBERT моделі [36], спеціально розробленої для моделювання репрезентацій фраз, значно нижча, ніж у загальних моделей моделювання подань речень (E5 base, GTE base, SimLM, Jina v1). Дійсно, показник точності (Таблиця 4.5) для PhraseBERT моделі найнижчий у порівнянні з моделями глибоких нейронних мереж як наведено вище. Отримані дані знаходяться у відповідності з попередніми результатами досліджень [38].

Порівнюючи результати застосування моделі PhraseBERT з моделлю неглибоких нейронних мереж (Fasttext [30]), доцільно зауважити, що PhraseBERT має нижчі показники точності. Це може бути пов’язано з різними масштабами попереднього тренування моделей. Дійсно, PhraseBERT навчали на синтетично згенерованих парафразах і фразах з контекстами, витягнутими з корпусу Book3, тоді як інші загальні моделі використовували різноманітні великомасштабні набори даних. Ці набори даних містили різні реальні дані, як-от інтернет-форуми (пари запитань-відповідей), соціальні мережі (заголовок-коротке резюме), наукові статті, бази знань і набори даних подібності речень, анотовані вручну (NLI, SNLI тощо). Це призвело до здатності моделей, навчених на цих даних, бути більш стійкими до зміни домену. Також має недоліки, які перераховані вище, модель McPhraSy, так як вона використовує подання з PhraseBERT без подальшого тренування, що

враховано та виправлено при теоретичному обґрунтуванні архітектури Phrase2Vec.

“GPT sentences aug” [71], навчений на синтетичних реченнях, які повинні містити певні навички, демонструє перспективність використання великих мовних моделей для генерації реалістичних даних і потенціал для вилучення знань з них. Однак навчання на реальних даних забезпечує більш різноманітне і нюансоване розуміння мови, охоплюючи складність і мінливість, притаманні реальним контекстам.

Перспективи подальшого дослідження мають базуватися на парадигмі Masked-Auto-Encoders для некерованого попереднього навчання подань речень [87], які змушують модель конденсувати значення речення в поданні спеціального токену. Адаптація такого підходу некерованого попереднього тренування може бути корисною для навчання подань фраз у контексті. Вивчення адаптації запропонованого підходу для багатомовних текстів є важливим, оскільки він відповідає зростаючому попиту на крос-культурність та різноманітні додатки в цій галузі.

Отримані дані показують позитивні результати щодо застосування архітектури Phrase2Vec в порівнянні з іншими моделями.

4.4 Сумаризація документів у сфері управління людськими ресурсами

Сумаризація є процесом автоматичного створення анотацій. В експериментах використано реальні резюме та вакансій з бази даних Daxtra Technologies. Задля пришвидшення генерації експериментальних даних, для створення анотацій було використано ВММ (GPT-4o). Валідаційна частина, що застосовувалася для тестування (1 000 резюме та 500 вакансій), була додатково перевірена та вдосконалена професійним рекрутером.

4.4.1 Скорочення довгих документів для подальшої сумаризації

Підхід до скорочення довгих документів, як описано в розділі 2.4 було протестовано з задачі сумаризації резюме, які перевершують ліміт у 1 024 токенів, притаманний BART large. Тестування проводилося на 350 резюме, анотації для яких були створені з застосуванням ВММ і, задля забезпечення високої якості даних, пройшла перевірку та коригування експертом.

Таблиця 4.6 — Оцінка якості методів скорочення тексту

Методи скорочення тексту	Якість (Rouge 1)
перші 1024 токени (“наївне” скорочення)	0,53
секційне скорочення	0,55
секційне скорочення за алгоритмом [74]	0,555
секційне скорочення по запропонованому алгоритму	0,57

“Наївне” скорочення документів є обробкою регламентованої кількості перших токенів. Таке скорочення призводить до найнижчої якості, бо модель втрачає доступ до частини потенційно важливого тексту чи навіть деяких секцій загалом.

Секційне скорочення являє собою скорочення в рамках кожної з секцій. Таке скорочення дозволяє досягти кращих результатів, бо гарантує, що всі секції будуть присутні у поданому моделі тексті, але теж призводить до втрати інформації.

Секційне скорочення за алгоритмом [74] передбачає застосування фільтрації стоп-слів та найбільш частотних слів в документі. Таке скорочення хоча й демонструє дещо кращі результати в порівнянні з секційним скороченням, поступається запропонованій методиці збереження ключових фраз. Це може бути пояснено тим, що при фільтрації стоп-слів тексти перестають бути зв'язними, і модель важче навчати на таких даних. Сучасні

мовні моделі, такі як BART, навчали на текстах, де стоп-слова не видалялися, що дозволило їм зберігати “розуміння” мовних структур і контексту. Так само, як і BERT, BART, на відміну від TF-IDF, де акцент робиться лише на окремих термінах, не базується на простих статистичних методах, а використовує глибоке розуміння контексту і взаємозв'язків між словами в тексті. Тому для таких моделей тексти без стоп-слів можуть бути менш ефективними, оскільки втрачається важлива інформація про зв'язки між словами і фразами.

4.4.2 Некероване попереднє навчання з використанням структури документів та зваженої функції втрат

Некероване попереднє навчання моделі застосовують для адаптації моделі до цільового домену [35].

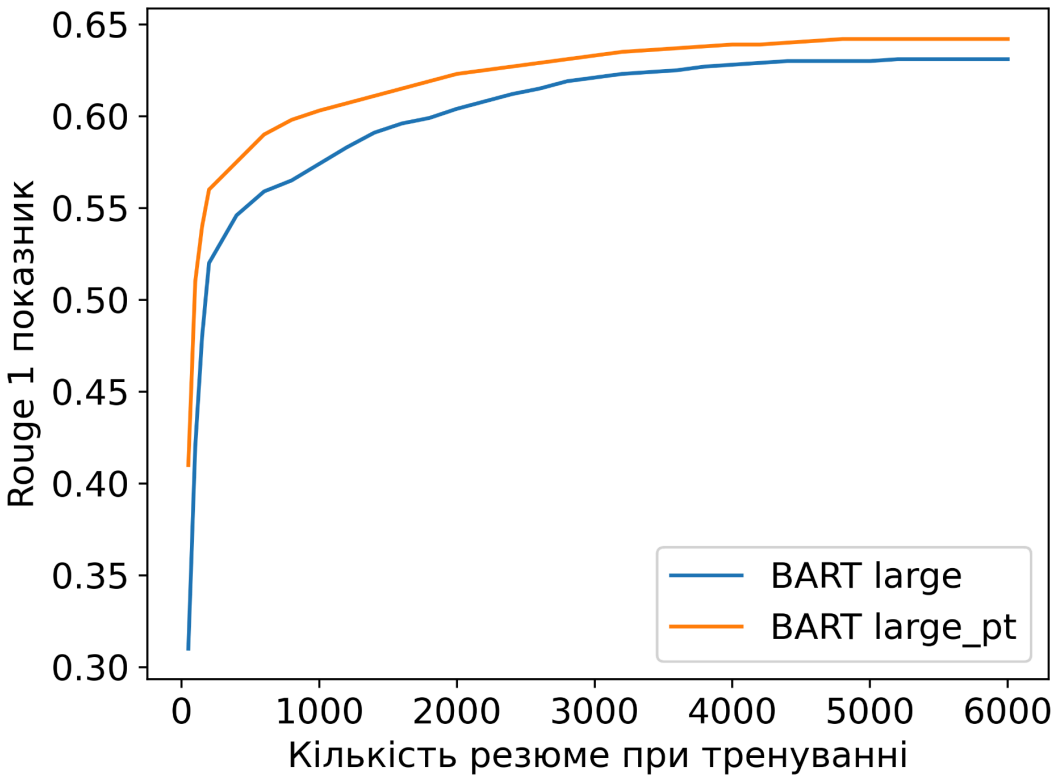


Рисунок 4.8 — Якість моделей при різних обсягах тренувальних даних

Як видно з рисунка 4.8, адаптація до домену з використанням функції втрат, орієнтованої на ключові фрази (BART_pt), є ефективною і забезпечує перевагу над продуктивністю базової моделі, що особливо помітно при невеликих обсягах навчального набору даних. Щодо масштабування обсягів навчального набору даних, можна спостерігати зменшення приросту ефективності при збільшенні кількості тренувальних зразків. Метрики перестають покращуватися після ~3 000 навчальних зразків. Це вказує на те, що модель досягла своєї здатності навчатися на доступних даних. Така поведінка свідчить, що після певного порогу — у випадку сумаризації резюме за допомогою GPT-4, це 3 000 зразків — додаткові навчальні дані тієї ж якості не призводять до покращення продуктивності, і доцільно обмежити розмір навчального набору даних до 3 000 зразків.

Таблиця 4.7 — Аналіз впливу попереднього тренування на якість сумаризації при різних обсягах кількості тренувальних зразків

Модель	Rouge 1	Rouge 2	Rouge L	NDCG
Розмір датасету = 50				
BART-large	0,31	0,135	0,241	0,48
BART-large_pt	0,41	0,187	0,315	0,68
% покращення	32%	38%	31%	42%
Розмір датасету = 1 000				
BART-large	0,574	0,328	0,436	0,75
BART-large_pt	0,603	0,356	0,466	0,82
% покращення	5%	8,5%	6,9%	9,3%
Розмір датасету = 6 000				
BART-large	0,631	0,382	0,495	0,81
BART-large_pt	0,645	0,390	0,501	0,84
% покращення	2,2%	2%	1,2%	3,7%

Як видно з таблиці 4.7, попереднє тренування з урахуванням ключових фраз є особливо корисним для задачі зіставлення вакансій і резюме, оскільки модель має тенденцію прогнозувати більш сфокусований і релевантний зміст. Це чітко проявляється у значному покращенні показників NDCG, особливо за невеликої кількості тренувальних зразків. Впровадження зважування токенів для обчислення втрат разом зі стратегією маскування, орієнтованою на ключові фрази, забезпечує підвищену увагу до токенів, пов’язаних із ключовими фразами. Це дозволяє моделі генерувати анотації, які є не лише точнішими, але й краще відповідають професійній і контекстуальній релевантності.

4.5 Подання текстів та документів у векторному просторі

4.5.1 Експериментальні дані

В експериментах щодо *некерованого подання текстів* у векторному просторі було використано великий та різноманітний набір даних, що складається приблизно з 4 мільйонів резюме та 1,4 мільйонів вакансій. Як резюме, так і вакансії були сегментовані відповідно до описаної в розділі 3.1.1 методики сегментації.

З набору даних, отриманого з резюме, було виділено 20,4 мільйона унікальних розділів із описом досвіду роботи. Після фільтрації записів з менше ніж шістьма реченнями, відформатованими в списковий формат, для аналізу було залишено 3,6 мільйона описів. Кожен вибраний розділ про досвід роботи був випадковим чином розділений на дві частини без повторень, при цьому кожен сегмент містив щонайменше три пункти.

Розділи, що визначають вимоги та бажаний профіль кандидату в вакансіях були попередньо оброблені таким же чином, як і розділи з описом досвіду роботи з резюме. Така попередня обробка дозволила створити 2,4 мільйона позитивних пар.

Розділи з резюме, що містили анотацію (summary), були виділені та перевірені на інформативність шляхом перевірки кількості фраз-навичок. Записи з неінформативними резюме, які не містили назви посади або принаймні 3 навичок, були відфільтровані. Розглядалися лише записи з резюме, які містили щонайменше 5 рядків у розділі про досвід роботи. Таким чином, було сформовано 996 тисяч унікальних пар “резюме — розділ про досвід роботи”.

Через дисбаланс даних, позитивні приклади з частин досвіду роботи резюме та частин вакансій були зменшені в кількості, щоб відповідати обсягу позитивних пар “анотація — розділ про досвід роботи”, забезпечуючи збалансоване навчання, обмежене двома ітераціями для пар “анотація — розділ про досвід роботи”.

Остаточна характеристика набору даних, використаного для некерованого навчання, наведена в таблиці 4.8.

Таблиця 4.8 — Показники набору даних, використаного для некерованого навчання

Джерело даних	Кількість зразків (тис)	Середня довжина (слів)
Пари “анотація — розділ про досвід роботи”, отримані з резюме (крос-секційне вирівнювання)	996	94
Частини розділу про досвід роботи (внутрішньосекційне вирівнювання)	996	61
Частини вакансії “вимоги” та “роль” (внутрішньосекційне вирівнювання)	996	48

Таким чином, зібраний датасет для некерованого навчання містить 2 988 тисяч позитивних пар текстів.

Для тренування *векторного подання документів* було створено синтетичний набір даних. Сто тисяч резюме, які містять щонайменше три розділи про працевлаштування, були використані для експериментів. Використовується найновіший розділ про працевлаштування. Статистика навчального набору даних наведена в Таблиці 4.9.

Таблиця 4.9 — Характеристики синтетичного набору даних, використаних для навчання модуля векторизації

Показник	Середня довжина (слів)
секція 1 (мета, резюме, навички)	146
секція 2 (2-ий та 3-ій описи останніх місць роботи)	148
секція 3 (освіта, тренування, сертифікації, публікації)	53
синтетична вакансія, згенерована BMM	232

Згенерований синтетичний набір даних доцільно охарактеризувати як придатний для подальших досліджень.

Оцінка людиною (рекрутером) є ключовою для *оцінювання систем рекомендацій резюме*. Оцінку запропонованої технології проводили за допомогою набору даних, складеного з реальних даних рекрутерів, наданих приватною онлайн-рекрутинговою компанією.

Для забезпечення високої якості даних застосовуються методи очищення та усунення упередженості. Зокрема, для виявлення мови використовується модель FastText з відкритим кодом, і записи з неангломовним контентом виключаються з набору даних.

Остаточний набір даних для оцінки складається з 12 068 резюме та 200 описів вакансій, що точно відображає динамічну та різноманітну природу реальних ринків праці. Для забезпечення стійкості та узагальнюваності, набір даних включає резюме та описи вакансій з різних галузей та професійних секторів.

Набір даних складається з двох версій — повнотекстових документів та анотацій, згенерованих за допомогою великої мовної моделі (ВММ). Ця подвійна репрезентація дозволяє здійснити більш ретельну оцінку моделей на різних рівнях деталізації та абстракції. Повнотекстова версія відображає реальні дані, з середньою довжиною 509 слів у вакансіях та 862 слова на резюме.

Натомість версія на основі анотацій пропонує значно коротші записи — в середньому 96 слів на вакансію та 102 слова на резюме, що робить її придатною для бенчмаркінгу класичних моделей векторизації тексту. Версія на основі анотацій зроблена публічно доступною для сприяння подальшим дослідженням у цій галузі [5].

Для оцінки запропонованої системи рекомендацій використовували косинусну подібність як фінальну метрику для рекомендації резюме до конкретної вакансії.

4.5.2 Практичне застосування та оцінка результатів

Некероване навчання подань текстів проводилося за допомогою множинного негативного ранжованого збитку, як визначено у рівнянні (2.2).

Урізання тексту для моделі HR Embedder була встановлене в розмірі 128 токенів, оскільки емпірично було виявлено, що подальше збільшення довжини послідовності не призводить до помітного покращення ефективності, водночас збільшуючи вимоги до обчислювальних ресурсів.

Ефективність моделі на валідаційному наборі реальних відповідних пар вакансій і резюме (на основі автоматично згенерованих анотацій) вимірювалася під час попереднього навчання HR Embedder.

Експерименти проводилися на GPU T4, для контрастивного навчання HR Embedder використовувався розмір батчу 64. Для оптимізації застосовувався оптимізатор AdamW, з коефіцієнтом швидкості навчання $2 \cdot 10^{-5}$ і лінійним розігрівом протягом перших 10% кроків навчання.

Некероване навчання *векторного подання документів (*модулю векторизації) на синтетичному наборі даних проводили з поданнями резюме

та синтетично згенерованими вакансіями, як визначено в рівнянні множинного негативного ранжованого збитку (2.2), за допомогою Множинного Негативного Рангового Збитку. Експерименти проводилися на GPU A100, розмір батчу був встановлений на 96. Використовувався оптимізатор AdamW з коефіцієнтом швидкості навчання $2 \cdot 10^{-5}$ і лінійним розігрівом, який застосовувався протягом перших 10% кроків навчання.

Для перевірки ефективності запропонованого *м*етоду некерованого навчання моделі HR Embedder був проведений порівняльний аналіз з існуючими сучасними підходами до векторизації загальних текстів, а також з спеціальними методами, розробленими для підбору відповідності резюме та вакансій — ConFit [78]. Оскільки ваги моделі ConFit відсутні у загальному доступі, модель “BERT-base” була донавчена на тих самих даних згідно зі стратегією, описаною в оригінальній статті, яка описує ConFit (EDA та доповнення ChatGPT).

Результати експериментів представлені в Таблиці 4.10.

Таблиця 4.10 — Показники якості моделей подань автоматично згенерованих анотацій

Модель	MAP	MRR	NDCG
TFIDF	0,39	0,61	0,64
BERT-base	0,31	0,56	0,61
e5-base-v2	0,55	0,75	0,77
gte-base	0,56	0,75	0,78
SimCSE*	0,421	0,642	0,688
DeCLUTR*	48,93	0,681	0,728
ConFit	0,59	0,775	0,79
HR Embedder	0,61	0,79	0,81

Оригінальний BERT без специфічного донавчання демонструє гірші результати, ніж TFIDF (MAP на 0,08 нижчий у порівнянні з TFIDF). “SimCSE” показує кращу продуктивність, ніж BERT, що підкреслює важливість адаптованих стратегій навчання. Але “SimCSE” поступається моделям, навченим під наглядом на даних, анотованих вручну.

“DeCLUTR” демонструє кращі результати порівняно з “SimCSE”. Це можна пояснити використанням різноманітних позитивних прикладів під час навчання шляхом вибору неперекривних текстових фрагментів з одного документа. Проте, як видно з Таблиці 4.10, випадковий вибір фрагментів, як це описано в стратегії “DeCLUTR”, без урахування внутрішньої структури резюме, не забезпечує достатньої семантичної узгодженості позитивних прикладів. Зокрема, отримані результати (покращення NDCG на 11%) підкреслюють важливість інтелектуального вибору позитивних пар, який враховує внутрішню структуру резюме.

Запропонований метод некерованого навчання моделі HR Embedder перевершує стратегію “ConFit” (покращення на 2,5%), яка покладається на аугментації, створені BMM, використовуючи виключно дані, отримані з резюме та оголошень про вакансії. Це робить його як більш економічно вигідним, так і масштабованим, оскільки він усуває залежність від обчислювально витратного створення синтетичних даних, водночас забезпечуючи високу продуктивність.

Результати обробки повнотекстових резюме та вакансій базовими моделями та розробленим модулем векторизації (*векторне подання документів)* наведені в таблиці 4.11.

Таблиця 4.11 — Показники якості векторизації при обробці документів

Метод	MAP	MRR	NDCG
TFIDF	0,42	0,62	0,66
BERT-base	0,19	0,39	0,47

e5-base-v2	0,47	0,71	0,72
gte-base	0,52	0,73	0,76
ConFit	0,60	0,79	0,80
Модуль векторизації	0,63	0,82	0,82

Порівняння даних таблиць 4.11 та 4.10, вказує на те, що використання резюме у їх повному текстовому вигляді призводить до суттєвого зниження продуктивності для загальних підходів на основі глибокого навчання (продуктивність “e5-base-v2” на 17% нижча при використанні повнотекстового резюме порівняно з продуктивністю, коли використовуються автоматично згенеровані анотації).

Запропонований модуль векторизації перевершує результати, отримані за допомогою базових моделей (таблиця 4.11).

Незважаючи на подібність передобробки документів (наявність сегментації в передобробці текстів для подальшого оброблення) з моделлю “ConFit”, представлений модуль векторизації дозволяє досягти покращення на 2% NDCG.

4.6 Крос-лінгвістична дистиляція векторних подань текстів

4.6.1 Експериментальні дані

Відповідно до дизайну MTEB для україномовних текстів, метою є забезпечення високої різноманітності зібраних даних з різних джерел, що дозволить моделювати широкий спектр реальних застосувань. Для вирішення цієї проблеми різноманітності слід розглянути різні домени, які б представляли як формальний, так і неформальний стилі. Крім того, слід забезпечити використання даних дозвільного характеру.

У контексті варіативності у середній довжині вибірки в різних наборах

даних, MTEB зазвичай включає дві різні версії наборів даних, що походять з одного джерела. Ці версії створюють з урахуванням необхідності забезпечити відмінності у довжині тексту, і одна версія зосереджена на порівнянні речень (версія з короткими текстами), а інша — на порівнянні абзаців (версія з довшими текстами). Зокрема, на прикладі даних, зібраних зі StackExchange, було скомпоновано два набора даних. Перший набір даних містить заголовки запитань, тоді як другий набір даних включає як заголовок, так і супровідний опис.

Джерело 1 для датасету: Законопроекти Верховної Ради України

Першим джерелом даних, обраним для аналізу, є Законопроекти Верховної Ради України. Цей набір даних було створено на основі вилучення назв законопроектів разом із пов'язаними з ними категоріями. Для порівняльного аналізу було відібрано дев'ятнадцять найпоширеніших категорій, зокрема ті, що стосуються комітетів, які займаються “правоохоронною діяльністю”, “фінансами, податковою та митною політикою” та іншими питаннями. В результаті цього процесу відбору було отримано корпус, що складається з 9 678 прикладів, середня довжина яких становить 26,1 слова.

Джерело 2 для датасету: Назви книг та їхні описи

Друге джерело даних охоплює назви книжок разом з їхніми описами та відповідними категоріями — загалом 12 590 зразків у 53 різних категоріях. Цей набір даних відрізняється від першого, оскільки зосереджується на літературних творах, що охоплюють широкий спектр категорій. Він охоплює широкий спектр тем і сюжетів, починаючи від технічних галузей, таких як “Комп'ютерна література”, до досліджень людської психології та стосунків у “Психологія та взаємини”, а також творчих і фантастичних оповідань, призначених для дітей. Кожний приклад в цьому наборі даних складається в середньому зі 126,8 слів, що дає змогу перевірити здатність моделей обробляти довші тексти.

Джерело 3 для датасету: Петиції до українського уряду

Третім джерелом даних є веб-сайт, призначений для розміщення петицій. Ця платформа слугує цифровим простором, де люди можуть створювати та підписувати петиції з різних соціальних, політичних та екологічних питань. На відміну від попередніх наборів даних, зосереджених на законодавчих документах і літературних творах, це джерело відображає громадські настрої та активність, надаючи уявлення про суспільні проблеми та адвокаційні зусилля. Цей набір даних охоплює колекцію з 3 387 зразків, розподілених за 18 окремими категоріями. Завдяки об'єднанню петицій, цей набір даних пропонує різноманітний спектр тем, що відображає безліч інтересів та справ, які відстоюють онлайн-спільноти. Аналіз цих даних дозволяє дослідникам вивчати тенденції громадської думки, а також оцінювати здатність моделей коректно представляти тексти в стилі суспільного дискурсу та політичного порядку денного.

Джерело 4 для датасету: Український форум

Четвертим джерелом даних, обраним для аналізу, є форум української спільноти. Це джерело даних було обране для того, щоб зафіксувати автентичне використання розмовної мови та охопити широкий спектр тем. Цей набір даних має на меті надати уявлення про реальні моделі спілкування, що відображають розмовну мову та неформальний дискурс, які часто зустрічаються на форумах онлайн-спільнот. Оскільки користувачі ставлять запитання не лише українською, а й іншими слов'янськими мовами, для фільтрації неукраїномовних даних була використана мовна модель FastText. Набір даних містить 86 різних категорій, серед яких “Стиль, Мода, Бренди > Бренди”, “Сім’я, Дім, Діти > Домашні тварини” тощо.

Як і в реальних текстах, в зразках цього набору трапляються граматичні помилки, друкарські помилки та суржик (суміш української та інших мов, поширених у регіоні).

Джерело 5 для датасету: Український портал новин

Набір даних, зібраний для цього дослідження, складається з українських новинних статей, отриманих з найбільших українських новинних

агентств. Таке доповнення до джерел даних має на меті забезпечити комплексний погляд на використання мови в журналістиці та медіасфері. Новинні статті пропонують окремий погляд на використання мови, що характеризується формальними структурами, журналістськими конвенціями та спеціалізованою термінологією.

Цей набір даних містить 23 категорії, серед яких “війна”, “бокс”.

Приклади з наборів даних надано в Таблиці 4.12

Таблиця 4.12 — Приклади зібраних даних

Джерело даних	Приклад	Категорія прикладу
Законопроекти Верховної Ради України	Проєкт Закону про внесення змін до Закону України “Про адвокатуру та адвокатську діяльність” щодо вдосконалення окремих питань організації адвокатської діяльності	Правова політика
Назви книг та їхні описи	Проєкт Закону про внесення змін до Закону України “Про адвокатуру та адвокатську діяльність” щодо вдосконалення окремих питань організації адвокатської діяльності	Кулінарія. Їжа та напої
Назви книг та їхні описи	“Легкі й швидкі рецепти неперевершених десертів для святкового дня і просто гарного настрою! Ніжний торт, ароматний лимонний чізкейк, рум'яні кексики, пиріжки, еклери…”	Комунальне господарство
Петиції до українського уряду	“Про порядок повірок квартирних/будинкових лічильників води”	Бізнес, Фінанси, Нерухомість
Український форум	“Київ, 2-кімнатна в новобудові – по чому?”	Бізнес, Фінанси, Нерухомість
Український портал новин	“День фізичної культури і спорту в Україні: листівки та побажання”	Свята

Загальні статистичні характеристики еталонних наборів даних надано в таблиці 4.13.

Таблиця 4.13 — Статистичні характеристики наборів даних

Назва датасету	Кількість зразків у вибірці	Кількість класів у вибірці	Кількість категорій	Кількість зразків	Середня довжина зразку (букв)	Середня довжина зразку (слів)
UkrLawDrafts	9687	19	19	9 687	186,8	26,1
UrkBook Blurs	1250 - 1693	11 - 15	48	16 678	748,4	123,7
UkrPetitions (s2s*)	3474	18	18	3 474	75,75	10,8
UkrPetitions (p2p**)	3474	18	18	3 474	1332	197
UkrQAForum (s2s)	4380 - 8475	11 - 15	86	155 788	42,3	7,9
UkrQAForum (p2p)	4380 - 8375	11 - 15	86	155 781	311	56,9
UkrNews (s2s)	8159 - 8,457	15	23	74 746	72,3	12,3
UkrNews (p2p)	8159 - 8457	15	23	74746	152,4	25,4

Примітка:

* s2s означає, що в якості зразків беруться лише речення заголовка

**p2p — відповідно використовуються об'єднані абзаци заголовка та опису.

Для *тренування* моделі-студента подань українського тексту потрібні пари текстових зразків, узгоджені з багаторесурсною мовою (у нашому випадку — англійською). Навчальні дані зібрані з сайту Opus [80] і складаються з комбінації декількох корпусів. Зважаючи на неоднорідну якість даних та мовне забруднення, були застосовані кроки для первинної обробки текстів.

Конкретно тексти очищуються за допомогою наступних кроків:

— Дедуплікація;

— Ідентифікація мови та фільтрація зразків, що містять неукраїнські тексти;

— Відсіювання вибірок, що відрізняються за довжиною більш ніж удвічі.

Статистичні дані об'єднаного набору даних наведено в Таблиці 4.14.

Таблиця 4.14 — Статистика об'єднаного набору даних, використаних для навчання моделі

Джерело даних	Кількість пар текстових зразків	Кількість унікальних текстових зразків
OpenSubtitles2018	877 780	419 004
Wikimedia	757 910	617 809
SciPar_Ukraine	306 813	306 813

Продовження таблиці 4.14

QED	215 530	198 100
TED2020	208 141	198 876
Tatoeba	175 502	168 480
ELRC-5179-acts_Ukrainian	129 942	126 361
ELRC-5180-Official_Parliament_Ukraine_Ukrainian_laws_EN	116 260	115 494
ELRC-5181-Official_Parliament_Ukraine_abstracts_UK_laws	61 012	60 885
ELRC-5174-French_Polish_Ukraine	36 228	35 597
Загалом	2 885 118	2 264 340 Глобально унікальних: 2 202 117

Після остаточного видалення дублікатів отриманий набір даних містить 2 202 тисяч пар текстових зразків.

4.6.2 Практичне застосування і оцінка результатів

Для мінімізації необхідних обчислювальних ресурсів ваги української мовної моделі було ініціалізовано на основі моделі RoBERTa-base [82], навченої на українських текстових даних.

Оптимізатор AdamW використовується з лінійним прогріванням

протягом 10% кроків. Коефіцієнт швидкості навчання встановлено на рівні $2 \cdot 10^{-5}$.

Для порівняльної оцінки крос-лінгвістичної дистиляції векторних подань україномовних текстів використовували три загальнодоступні моделі, що підтримують українську мову: LaBSE, distiluse-base-multilingual-cased, E5.

Крос-лінгвістичну дистиляцію векторних подань проводили у двох сценаріях:

— Ukr-distil_mse — функція втрат на основі середньоквадратичної помилки;

— Ukr-distil_cos — функція втрат на основі косинусної подібності.

В якості метрики для оцінювання використовували показник нормалізованої взаємної інформації (NMI), який є гармонійним середнім між однорідністю та повнотою. Значення цього показника може варіюватися від 0 до 1, де 1 означає ідеально повне маркування зразків. Ця метрика не залежить від абсолютних значень міток, і перестановка значень міток класу або кластера не впливає на результат.

Результати оцінювання представлені в Таблиці 4.15.

Таблиця 4.15 — Результати оцінювання (NMI метрика) на еталонних датасетах

	LaBSE	distiluse	E5	Ukr-distil_mse	Ukr-distil_cos
UkrLawDrafts	0,175	0,138	0,201	0,145	0,234
UrkBookBlurs	0,513	0,410	0,481	0,329	0,542
UkrPetitions (s2s)	0,138	0,162	0,205	0,122	0,246
UkrPetitions (p2p)	0,249	0,222	0,274	0,121	0,326
UkrQAForum (s2s)	0,188	0,144	0,300	0,133	0,326

Продовження таблиці 4.15

UkrQAForum (p2p)	0,396	0,326	0,459	0,207	0,453
UkrNews (s2s)	0,559	0,416	0,627	0,517	0,568
UkrNews (p2p)	0,585	0,485	0,644	0,602	0,621
Середній показник	0,350	0,289	0,40	0,272	0,414

Як видно з результатів, використання функції втрат косинусної подібності має явну перевагу над середньоквадратичною помилкою (0,414 проти 0,272 середнього показника, що становить покращення абсолютного значення NMI на 14,2%). Тому доцільно рекомендувати саме функцію косинусної подібності при крос-лінгвістичній дистиляції текстових подань (Ukr-distil_cos).

4.6.3 Залежність якості отриманих текстових подань від моделі-вчителя

Були проведені експерименти з використанням різних моделей-вчителів. На рисунку 4.9 візуалізовано розподіли косинусної подібності, обчисленої між усіма можливими парами векторів з 5 000 випадкових англійських речень із навчального набору. Як видно, модель E5 демонструє дуже вузький розподіл.

Рисунок 4.9 показує кумулятивний розподіл оцінок косинусної подібності, де вісь *x* представляє поріг косинусної подібності, а вісь *y* показує відсоток записів, що мають подібность, меншу за цей поріг.

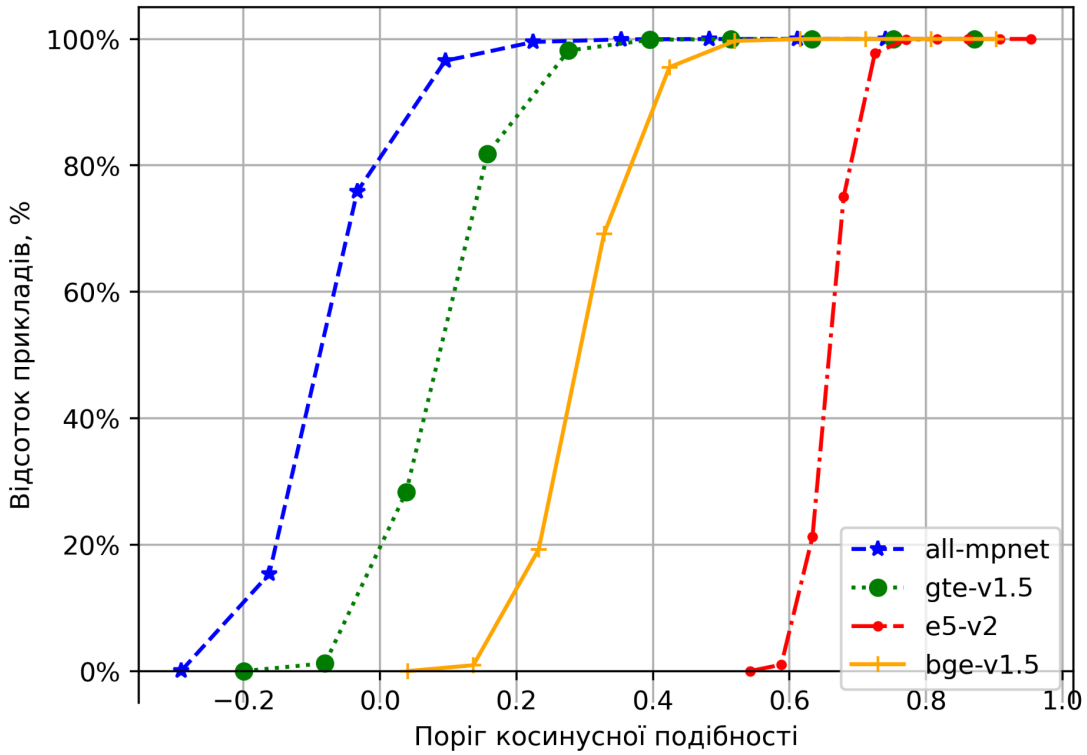


Рисунок 4.9 — Кумулятивні розподіли оцінок подібності між випадковими парами речень

Кореляція між розподілом оцінок косинусної подібності моделі-вчителя та продуктивністю на прикладних завданнях моделі-студента, отриманої шляхом дистиляції, наведена в таблиці 4.16.

Таблиця 4.16 — Кореляція між розподілом оцінок косинусної подібності та якістю дистиляції векторних подань текстів

	e5 base	bge base	gte base	all-mpnet
UkrLawDrafts	0,062	0,113	0,128	0,234
UrkBookBlurs	0,245	0,342	0,364	0,542
UkrPetitions (s2s)	0,054	0,124	0,105	0,246

Продовження таблиці 4.16

UkrPetitions (p2p)	0,045	0,112	0,199	0,326
UkrQAForum (s2s)	0,065	0,133	0,132	0,326
UkrQAForum (p2p)	0,127	0,215	0,263	0,453
UkrNews (s2s)	0,308	0,516	0,448	0,568
UkrNews (p2p)	0,414	0,601	0,548	0,621
Середній показник	0,165	0,269	0,273	0,414
90-ий перцентиль косинусної подібності	0,75	0,48	0,31	0,16

Як видно з таблиці 4.16, існує сильна кореляція між розподілом косинусної подібності між векторами нерелевантних речень та якістю текстових подань, отриманих моделлю-студентом. Кореляція Пірсона між “90-м процентилем косинусної подібності” та “середнім показником” становить -0,96, що інтерпретується як сильна негативна кореляція.

Низька подібність між нерелевантними векторами забезпечує більш стабільне навчання моделі-студента та формування більш розрізнених текстових подань. Це, у свою чергу, покращує здатність моделі точно захоплювати семантичні нюанси тексту.

Такі подання забезпечують кращу репрезентацію основної інформації та здатність розрізняти різні типи контенту, що призводить до підвищення ефективності виконання прикладних завдань.

Отже, вибір моделі-вчителя з нижчим середнім показником косинусної подібності між не пов’язаними між собою (випадковими) текстами є ключовим для ефективного перенесення знань та загального успіху моделі-студента.

4.7 Дистиляція моделей

За базову модель прийнято модель розміру BERT-base з 108 мільйонами параметрів, в якості моделі-вчителя використано модель розміру RoBERTa-large з 355 мільйонами параметрів, в якості цільової — MiniLMv2 з архітектурою L12xH384 з 41 мільйонами параметрів.

Задачі сегментації та парсингу ключової інформації вирішуємо як завдання класифікації токенів. Для полегшення оцінювання проводиться дослідження усереднених значень метрик.

При тренування на маркованих людиною даних базова модель досягла середніх показників якості 86,8 F1 метрика, модель-вчитель — 89,2 F1 метрика, цільова модель без використання методу дистиляції була здатна досягти 83,4.

На рисунку 4.10 показано процес дистиляції моделі-вчитель в модель-студента цільового розміра (цільову).

При процесі дистиляції 55 мільйонів токенів достатньо в задачі класифікації токенів для цільової моделі для досягнення якості на рівні базової моделі (рисунок 4.10). Подальше збільшення обсягу не маркованих даних для дистиляції дозволяє досягти перевершення якості базової моделі на 0,5% в абсолютних показниках (показник якості цільової моделі з застосуванням методу дистиляції — 87,3). При цьому цільова модель в 2,6 рази менша за базову.

Можна спостерігати зменшення приросту якості при збільшенні кількості токенів в тренувальному наборі. Метрики перестають покращуватися після ~120 мільйонів токенів в дистиляційному наборі даних.

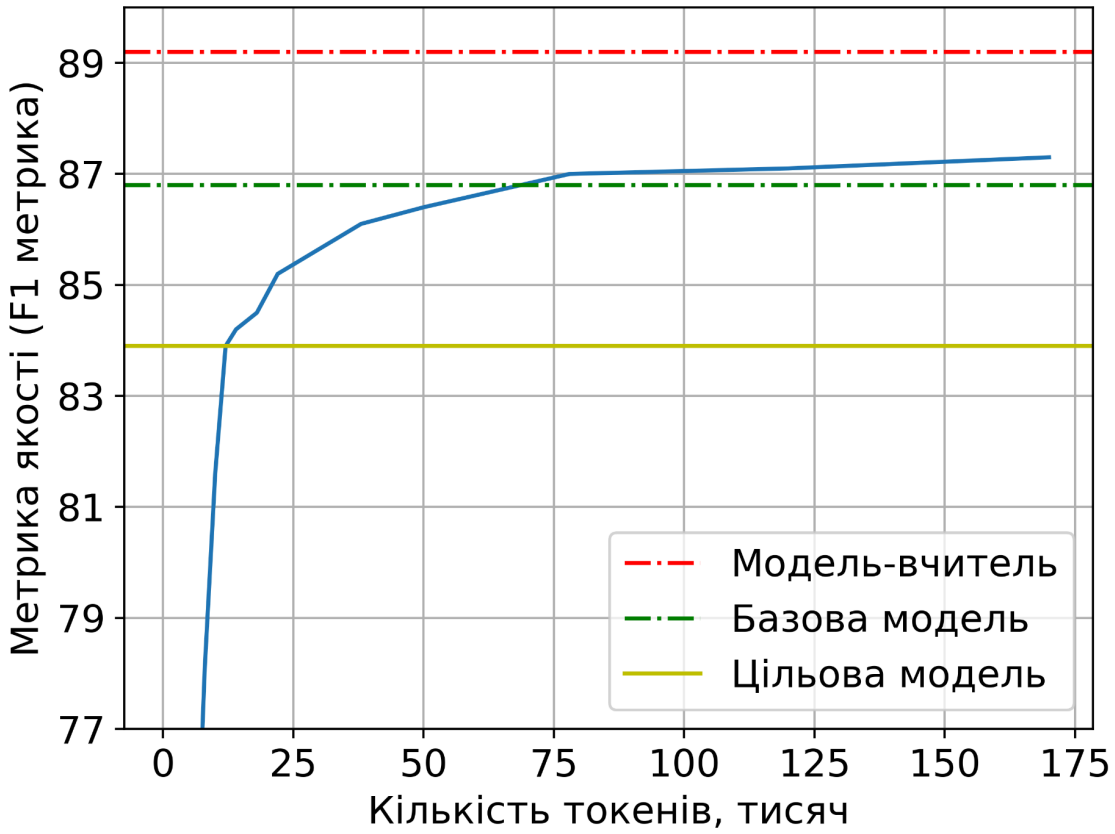


Рисунок 4.10 — Залежність метрик якості дистиляції від обсягу даних при дистиляції моделі-вчителя в модель-студента

Зупинка покращення метрик вказує на те, що модель досягла своєї здатності навчатися моделювати розподіл вірогідностей моделі-вчителя. Така поведінка свідчить, що після певного порогу — у випадку дистиляції моделі розміру 335 мільйонів параметрів у модель розміром 41 тисяч параметрів, це 120 мільйонів токенів — додаткові дані не призводять до покращення продуктивності, і доцільно обмежити розмір набору даних використовуваних для дистиляції до 120 мільйонів токенів.

Висновки за розділом 4

1. Встановлено, що застосування безпосередньої інтеграції параметрів про стильові ознаки (без використання комп’ютерного зору) дозволяє досягти

покращення якості класифікації токенів в завданні сегментації резюме на 3-6% F1 метрики у порівнянні з адаптованою до домену версією моделі BERT, навченою за допомогою MLM стратегії навчання. В завданні виділення ключової інформації (такої як назви компаній, фраз-навичок, назв посад тощо) якість класифікації токенів було підвищено на 2-10% F1 у порівнянні з адаптованою до домену версією моделі BERT, навченою за допомогою MLM стратегії навчання.

Встановлено, що безпосереднє включення інформації про стиль і форматування є корисною адаптацією оригінальної версії BERT щодо контекстуалізованого подання токенів у векторному просторі для візуально насичених документів.

В порівнянні з архітектурою, яка використовує комп'ютерний зір, розроблена модель дозволяє досягти 1,67 пришвидшення обробки документів.

2. Запропонований підхід — VacancySBERT, який використовує стратегію напівкерованого навчання для моделювання подань назв посад на основі навичок, зазначених у їхньому описі, дозволяє досягти більш точної нормалізації назв посад, оскільки дозволяє ідентифікувати схожі назви посад на основі необхідних навичок, які зазначені в описі роботи, навіть якщо самі назви посад відрізняються. Цей підхід дозволяє подолати необхідность в ручному маркуванні даних, що може бути дорогим і тривалим процесом, особливо з урахуванням швидких змін на ринку праці.

Переваги запропонованого підходу можна побачити у значному покращенні семантичного пошуку текстової подібності в задачі нормалізації назв посад. Дійсно, новий підхід перевершує сучасні енкодери для семантичної текстової подібності у завданні нормалізації назв посад на 10-21,5%, залежно від використання навичок під час інференсу.

3. Новий метод навчання з контекстно-орієнтованим вирівнюванням подань фраз призводить до значного покращення якості подань фраз на 9,9-10,6% в залежності від конкретного сценарію.

Встановлено, що Phrase2Vec безперебійно працює з новими навичками, які не були присутні в тренувальному наборі даних. Новий еталон семантичної подібності для навичок, витягнутих з реальних текстів з управління персоналом, разом з відповідними контекстними абзацами дозволяє надійно порівнювати моделі подання фраз навичок.

Встановлено, що розроблена структурна модель представлення фраз у контексті може бути використана в реальних резюме та вакансіях, дозволяючи врахувати нюанси порівняння фраз навичок.

Встановлено, що підхід “всі негативні пари” при тренуванні в умовах асиметричного датасету при використанні функції множинних негативних втрат при ранжуванні недоцільний для використання та призводить до зниження метрик.

4. Скорочення тексту секцій проходячи речення за реченням та зберігання лише ключові фраз до досягнення бажаної довжини дозволяє досягти 5% покращення якості сумаризації для довгих документів в домені управління людськими ресурсами в порівнянні з класичними методами скорочення та методикою видалення стоп слів та найбільш частотних слів.

Використання структури документів та зваженої функції втрат з використання коефіцієнтів вагомості для токенів, які формують ключові фрази, дозволяє досягти 10% покращення при подальшому тренуванні на цільових даних. Особливо корисною є така стратегія попереднього навчання, коли кількість тренувальних зразків обмежена.

Встановлено поріг у 3 000 зразків для сумаризації резюме за допомогою згенерованих GPT-4 анотацій, після якого додаткові навчальні дані тієї ж якості не призводять до покращення продуктивності, і доцільно обмежити розмір навчального набору даних до 3 000 зразків.

5. Для оцінки ефективності моделей у завданні зіставлення резюме та оголошень про роботу був зібраний еталонний набір даних. Цей набір складається з 200 вакансій та 12 068 резюме в двох форматах — повнотекстова версія та скорочена версія, згенерована великою мовною моделлю (ВММ).

Розроблена стратегія некерованого навчання для текстів в домені управління людськими ресурсами на основі внутрішньо- та міжсекційного вирівнювання, що призвело до покращення NDCG на 11 % порівняно зі стратегією DeCLUTR (тестування проводилося на скороченій версії набору даних).

Ця технологія забезпечує лінійну часову складність для зіставлення оголошень про роботу та резюме, а також покращення NDCG на 2% та 6% порівняно зі стратегією “ConFit” без та з фільтруванням результатів згідно з зазначеними у вакансії жорстким вимогам відповідно (повнотекстова версія набору даних).

6. Показано, що використання функції втрат на основі косинусної подібності призводить до значних покращень (на 14,2% за абсолютним показником NMI) у порівнянні з використанням функції втрат середньоквадратичної похибки при дистиляції векторних подань текстів з добре натренованої моделі-вчителя.

Доведено, що вибір моделі-вчителя та розподіл косинусної подібності, який вона генерує, впливають на якість дистиляції векторних подань моделі-студента. Встановлено, що існує сильна негативна кореляція між “90-м перцентилем розподілу косинусних коефіцієнтів подібності” та “середнім показником NMI”, отриманим на кластеризаційному еталоні.

Представлено новий і перший у своєму роді еталон для тестування українських текстових подань, який охоплює 5 різних доменів.

Модель та еталон для українських текстових подань доступні за посиланням: https://github.com/maiiabocharova/UkrMTEB.

7. Показана доцільність застосування процесу дистиляції задля пришвидшення моделей. Було показано, що при використанні моделі-вчителя з кількістю параметрів 355 мільйонів, модель-студент перевершує показники якості базової моделі (108 мільйонів параметрів) при обсягу дистиляційних даних в 55 мільйонів токенів. Встановлено, що 120 мільйонів токенів — поріг обсягу даних для дистиляції, подальше збільшення розміру датасету після якого недоцільне.

5 ОЦІНКА ЕФЕКТИВНОСТІ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ОБРОБКИ ПРИРОДНОМОВНИХ ТЕКСТІВ

5.1 Оцінювання інформаційної технології інтенсифікації процесу відбору та ранжування резюме рекрутером

Як показано у розділі 3, інформаційну технологію інтенсифікації відбору та ранжування резюме рекрутером доцільно реалізовувати шляхом дистиляції кожного етапу технологічної схеми. Ця технологія може включати етап сумаризації або здійснюватися без нього.

Для *технології з сумаризацією:* швидкість технології "AI ResJobFit" складає 0,45 секунди для резюме та 0,37 для кожної вакансі при використанні графічної карти T4 та опрацюванні документів з використанням батчингу; для *технології без сумаризації*: швидкість технології "AI ResJobFit" складає 0,29 секунди для резюме та 0,22 для кожної вакансій при використанні графічної карти T4 та опрацюванні документів з використанням батчингу. Такі результати не впливають на швидкість рекрутинга, так як проводяться одноразово (під час індексації документів).

При використанні технології без сумаризації, рекрутер має ознайомлюватися з повним резюме, а при включенні сумаризації у технологічну схему — лише з анотацією резюме. Час опрацювання резюме та анотації кожним рекрутером надано на рис 5.1 та 5.2 відповідно.

Для оцінювання середнього часу на ознайомлення з документами було використано методику, описану в розділі 2.3.

Як можна побачити, більшість відповідей припадає на часовий проміжок 6-10 секунд, при чому середній показник становить 7,9 секунд, що є у відповідності до попередніх досліджень [88]

На рисунку 5.2 надано розподіл середнього часу, витраченого рекрутером на ознайомлення з автоматично згенерованою анотацією резюме.

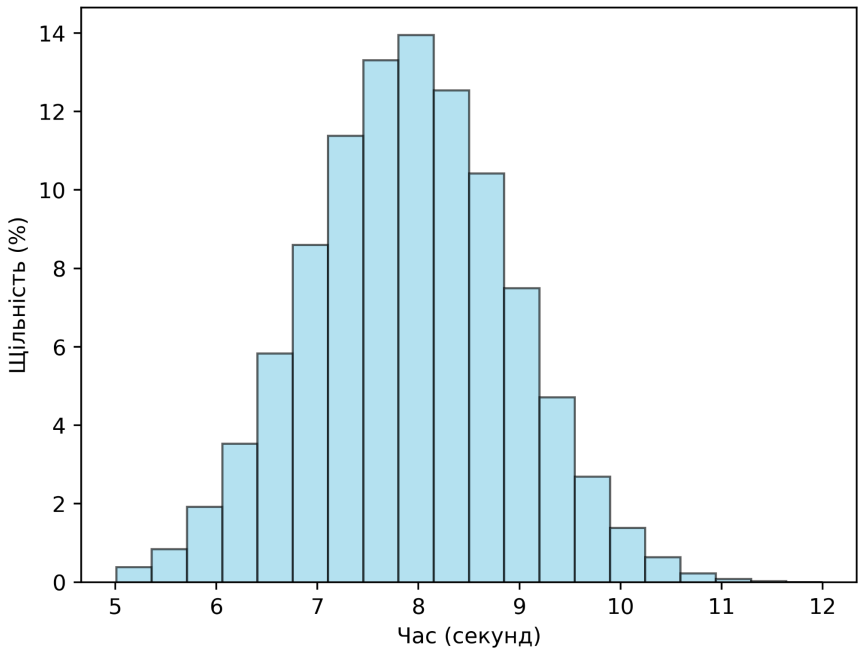


Рисунок 5.1 — Усереднений час, витрачений на ознайомлення з одним резюме рекрутером

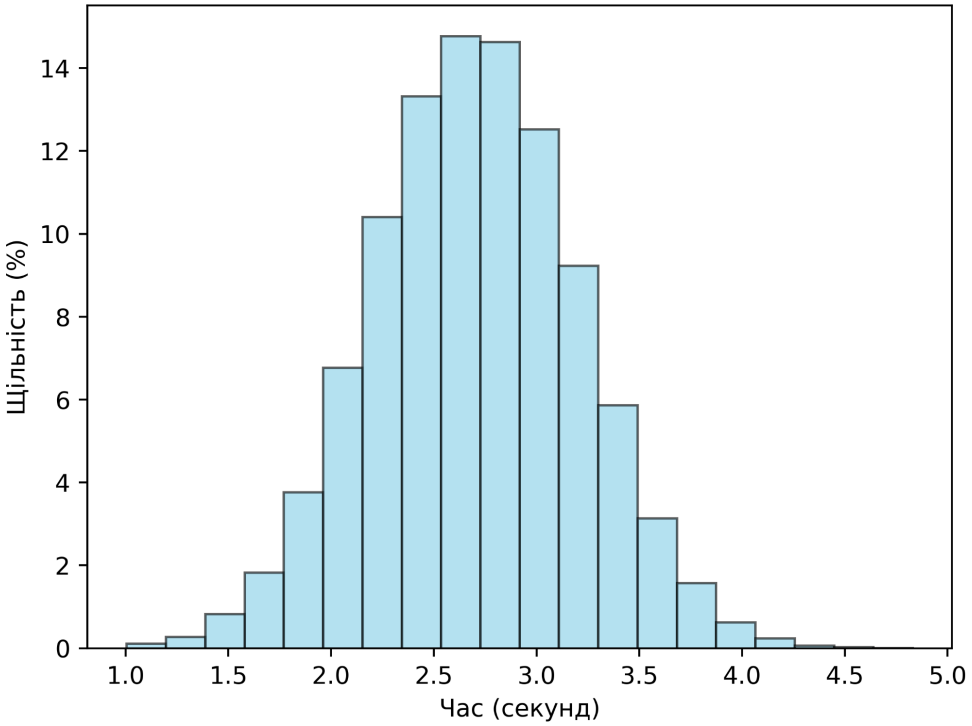


Рисунок 5.2 — Усереднений час, витрачений рекрутером на ознайомлення з автоматично згенерованою анотацією резюме

Швидкість ознайомлення з резюме при використанні структурованих анотацій зросла у 3,2 рази у порівнянні зі сценарієм відсутності анотацій (2-3,5 секунди на ознайомлення, при чому середній час становить 2,7 секунди).

Усереднення результатів, наданих на рис 5.1 та 5.2 дозволяють стверджувати про значне підвищення швидкості ознайомлення з документом рекрутером при використанні технології з сумаризацією. Порівняльний аналіз показників ефективності процесу надано в таблиці 5.1

Таблиця 5.1 — Порівняльний аналіз показників ефективності технології інтенсифікації процесу відбору та ранжування резюме рекрутером

Характеристика	Якість запропонованого рекрутеру матеріала, NDCG	Час ознайомлення рекрутером з одним документом, с
Технологія з сумаризацією	0,860	2,7
Технологія без сумаризації	0,845	7,9

Як видно з даних таблиці 5.1, швидкість ознайомлення рекрутера з одним документом для варіанта технології з використанням етапу сумаризації підвищується у 2,9 разів порівняно із варіантом технології без сумаризації. Дійсно, сумаризація дозволяє рекрутеру ознайомлюватися вже із згенерованою анотацію, а не з повним резюме. Це вказує на значну інтенсифікацію процеса рекрутинга і дозволяє рекомендувати технологію інтенсифікації процесу відбору та ранжування резюме рекрутером у варіанті технології з сумаризацією (рисунок 5.3).

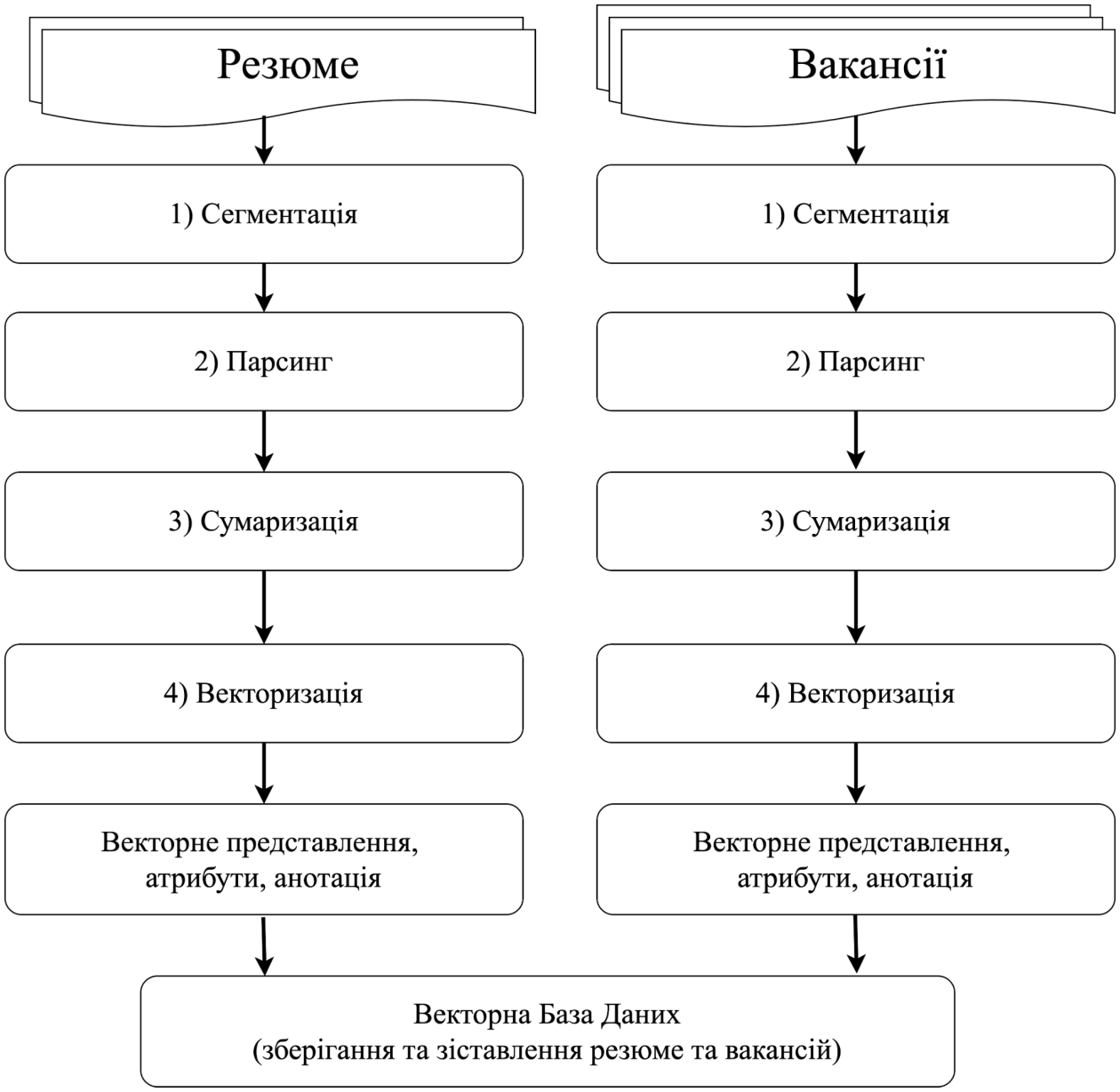


Рисунок 5.3 — Інформаційна технологія інтенсифікації процесу відбору та ранжування резюме рекрутером

Таким чином, документом, з яким ознайомлюється рекрутер у запропонованій технології є анотація резюме, що дозволяє підвищити ефективність рекрутингу.

5.2. Оцінювання технології аналізу резюме та зіставлення з вимогами вакансій в умовах відсутності рекрутера

Технологію аналізу резюме та зіставлення з вимогами вакансій в умовах відсутності рекрутера доцільно проводити із використанням дистиляції кожного з етапів технологічної схеми, і така технологія може бути проведена з застосуванням етапу сумаризації та без нього. Порівняльний аналіз показників ефективності процесу, кожний з етапів технологічної схеми якого піддавали дистиляції надано у таблиці 5.2

Таблиця 5.2 — Порівняльний аналіз показників ефективності аналізу резюме в умовах відсутності рекрутера

Характеристика	Якість, NDCG	час аналізу 1 резюме, с	час аналізу 1 вакансії, с
Технологія з сумаризацією	0,860	0,45	0,37
Технологія без сумаризації	0,845	0,29	0,22

Збільшення швидкості технології аналізу резюме та зіставлення з вимогами вакансій резюме (в умовах відсутності рекрутера) без використання сумаризації на 0,16 секунди для резюме та 0,15 для вакансії порівняно з технологією з використанням сумаризації пояснюється відсутністю необхідності використовувати додаткову енкодер-декодер модель для генерації анотації. Якість технології без сумаризації знижується несуттєво порівняно з технологією із сумаризацією (в умовах без рекрутера), що

передбачає доцільність запропонувати технологію рекомендацій резюме в умовах відсутності рекрутера у варіанті без сумаризації (рисунок 5.4)

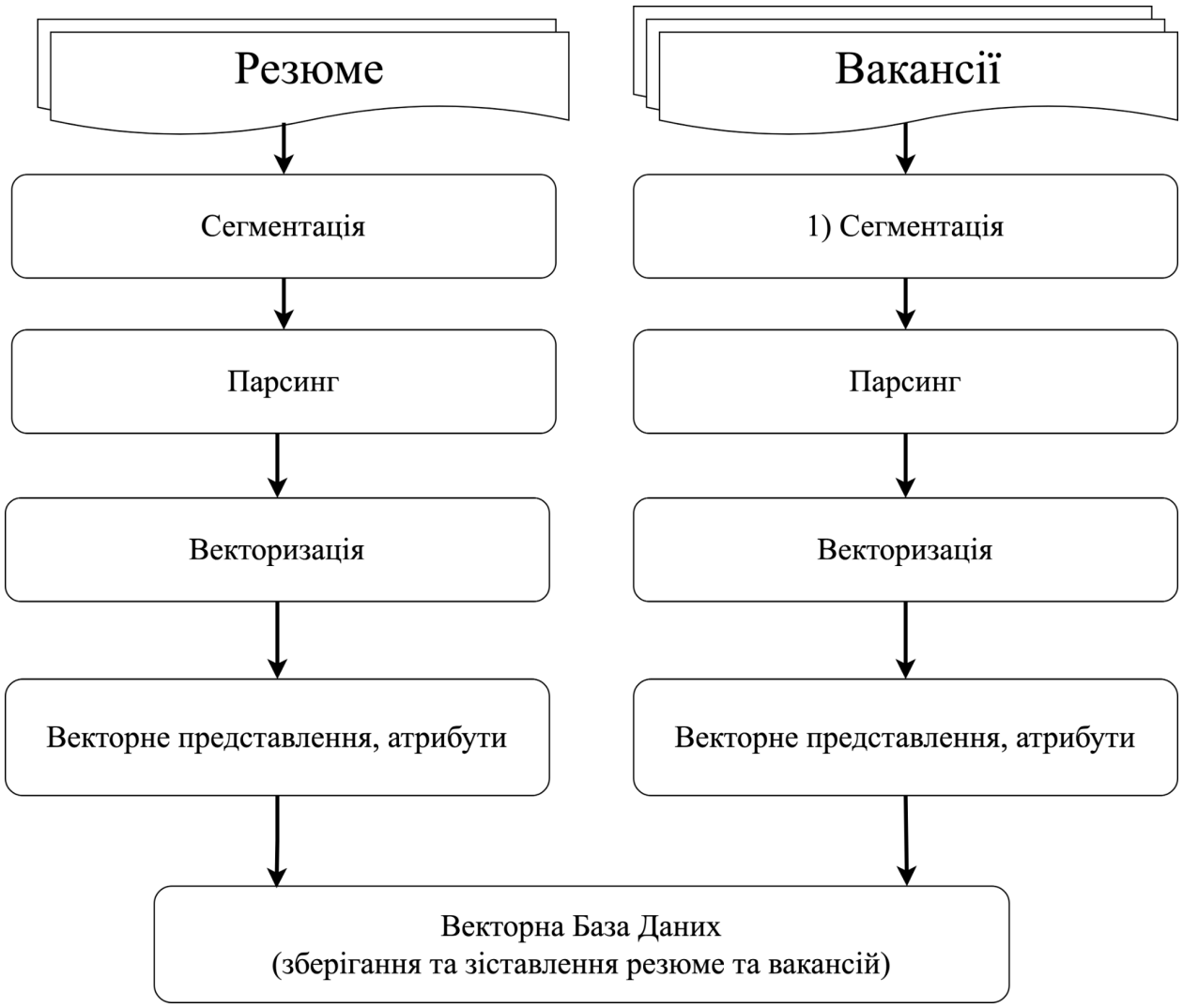


Рисунок 5.4 — Схема інформаційної технології аналізу резюме та зіставлення з вимогами вакансій в умовах відсутності рекрутера

Таким чином, в умовах відсутності рекрутера, етап сумаризації є недоцільним для застосування в інформаційній технології щодо аналізу резюме та зіставлення з вимогами вакансії.

5.3 Вплив окремих етапів на швидкість та якість інформаційної технології

Показники якості та швидкості інформаційної технології AI ResJobFit для кожного технологічного етапу було порівняно з даними, отриманими при застосуванні базової моделі.

5.3.1 Сегментація

Для оцінювання показників якості та відносної швидкості технології на етапі сегментації було використано 1 500 резюме. За базову модель прийнято BERT-base (модель без комп'ютерного зору), яка є текстовою, як і розроблена технологія "AI ResJobFit". Також для порівняння параметрів було досліджено модель комп'ютерного зору (LayoutLM v3 base), що дало змогу наочно оцінити переваги текстової технології "AI ResJobFit" у забезпеченні якісних подань токенів для подальшої обробки порівняно з підходом комп'ютерного зору.

Порівняльні параметри швидкості та якості етапу сегментації для обраних моделей надані в таблиці 5.1.

Таблиця 5.1 — Порівняння параметрів швидкості та якості етапу сегментації досліджуваних моделей

Модель	Відносна швидкість (швидкість моделі / швидкість базової моделі)	Якість (F1 score)
базова (BERT-base)	1	0,91
з комп'ютерним зором (LayoutLM 3 base)	0,6	0,93

Продовження таблиці 5.1

AI ResJobFit (без дистиляції)	1	0,94
AI ResJobFit (з дистиляцією)	2,35	0,94

Система з використанням комп’ютерного зору обчислює в рази більше інформації, тому її швидкість значно нижча за цей показник для інших моделей. Але результати дослідження вказують, що розроблена технологія “AI ResJobFit” на етапі “сегментація” дозволяє отримати вищі показники якості при значних перевагах у швидкості.

Якість найнижча в базової моделі, тому що вона не має доступу до стильової інформації, притаманної візуально насиченим документам. У випадку використання моделі “AI ResJobFit”, показники якості такі ж як у моделі, що використовує комп’ютерний зір, бо вона має доступ до стильової інформації, яка передається в неї безпосередньо.

Застосування методики дистиляції з більшої моделі в меншу з використанням [56] на етапі сегментації дозволяє досягти 2,35-кратного пришвидшення (таблиця 5.1).

Таким чином, етап сегментації за запропонованою технологією дозволяє підвищити якість в порівнянні з базовою моделлю. Ця якість є вищою в порівнянні з моделі з комп’ютерним зором, а її швидкість при використанні методики дистиляції майже в 4 рази вище.

5.3.2 Парсинг

Цей етап складається з двох частин - виділення ключової інформації та її нормалізації.

Для оцінювання показників якості та відносної швидкості виділення

ключової інформації було використано 2 500 секцій резюме.

Порівняльні параметри швидкості та якості щодо виділення ключової інформації для обраних моделей надані в таблиці 5.2

Таблиця 5.2 — Порівняння параметрів швидкості та якості виділення ключової інформації

Модель	Відносна швидкість (швидкість моделі / швидкість базової моделі)	Якість (F1 score)
базова (BERT-base)	1	0,69
з комп'ютерним зором (LayoutLM 3 base)	0,6	0,72
AI ResJobFit (без дистиляції)	1	0,72
AI ResJobFit (з дистиляцією)	2,35	0,71

Нормалізацію певних її компонентів (дати, локації, …) проведено за допомогою правил (так званий підхід "rule-based"). Для нормалізації більш варіативних сутностей (назв посад, навичок, назв компаній тощо) необхідно використання методів глибокого навчання. За базову модель прийнято PhraseBERT (модель без контекстної освіченості). В якості досліджуваної модель, яка враховує контекст фраз було використано McPhraSy.

Порівняльні параметри швидкості та якості щодо нормалізації варіативних сутностей методами глибокого навчання для обраних моделей надані в таблиці 5.3

Таблиця 5.3 — Порівняння параметрів швидкості та якості нормалізація варіативних сутностей

Модель	Наявність тренування в цільовому домені	Відносна швидкість (швидкість моделі / швидкість базової моделі)	Якість на задачі класифікації (F1 score)	Якість для ранжування (Recall@ 1)
Базова — враховує контекст (McPhraSy)	-	1	0,75	0,25
Базова — враховує контекст (McPhraSy)	+	1	0,90	0,36
PhraseBERT (не враховує контекст)	-	6,25	0,642	0,2
PhraseBERT (не враховує контекст)	+	6,25	0,78	0,30
AI ResJobFit (без дистиляції)	+	5,76	0,93	0,41
AI ResJobFit (з дистиляцією)	+	13,4	0,92	0,4

Моделі, які враховують контекст для моделювання подань фраз,

обробляють більше токенів в порівнянні з моделями, які моделюють подання фраз ізольовано. McPhraSy показує набагато нижчі показники швидкості, бо використовує дві моделі і множинні прогони тексту з іншою маскою для кожної фрази в абзаці.

Якість найнижча в базової моделі, тому що вона не має доступу до контекстної інформації, яка важлива для розуміння ключових слів та фраз. У випадку використання моделі “AI ResJobFit”, показники якості кращі, ніж в моделі яка використовує контекст спираючись на дві моделі. Це пояснюється тим, що вона має безпосередній доступ до контексту та, на відміну від McPhraSy, при тренуванні всі ваги моделі зазнавали тонкого налаштування.

При застосуванні фільтрації з використанням нормалізованих сутностей якість зіставлення резюме з вимогами вакансій підвищується на 2% NDCG.

Застосування методики дистиляції з більшої моделі в меншу з використанням методики, запропонованої у роботі [56], на етапі парсингу дозволяє досягти 2,35-кратного пришвидшення (таблиця 5.3).

Таким чином, етап парсингу за запропонованою технологією дозволяє підвищити якість в порівнянні з базовою моделлю. Ця якість є на рівні з комп’ютерним зором, але її швидкість при використанні методики дистиляції майже в 4 рази вище. Нормалізація сутностей зазнала значного пришвидшення (в 13,4 рази в порівнянні з базовою технологією, яка використовує контекст для моделювання подань фраз, при цьому було досягнуто покращення якості на 3-14%).

5.3.3 Сумаризація

Для технології “AI ResJobFit” для підбору резюме із поєднанням машинного навчання та оптимізованої праці людини важливим етапом є сумаризація, тому що цей етап покликаний скоротити час, який витрачає рекрутер на ознайомлення з резюме.

Базовою моделлю доцільно обрати BART large, так як саме ця модель показала найкращі показники в задачі сумаризації резюме [72].

Порівняльні параметри швидкості та якості етапу сумаризації для обраних моделей надані в таблиці 5.4

Таблиця 5.4 — Порівняльні параметри швидкості та якості етапу сумаризації для обраних моделей

Модель	Відносна швидкість	Якість (Rouge1 score)
базова (BART large)	1	0,57
ВММ (Phi 3.5)	0,11	0,63
AI ResJobFit (без дистиляції)	1	0,65
AI ResJobFit (з дистиляцією)	2,9	0,64

Результати, показані в таблиці 5.4 дозволяють зробити висновок, що попереднє тренування BART моделі з модифікацією функції втрат з урахуванням вагомості токенів дозволяє досягти значного підвищення якості. Дистиляція дозволяє досягти значного підвищення швидкості без суттєвого зниження якості.

5.3.4 Векторизація

Для оцінювання показників якості та відносної швидкості технології на етапі векторизації було використано 200 вакансій та 12068 резюме. За базову модель прийнято ConFit, яка є спеціально розробленою для текстів в домені управління людськими ресурсами, що відповідає цільовому домену

розробленою технології “AI ResJobFit”.

Порівняльні параметри швидкості та якості етапу векторизації для обраних моделей надані в таблиці 5.5.

Таблиця 5.5 — Порівняльні параметри швидкості та якості етапу векторизації для обраних моделей

Модель	Наявність дистиляції моделей на етапі векторизації	Використання згенерованої анотації	Відносна швидкість	Якість (NDCG)
базова (ConFit)	-	-	1	0,80
AI ResJobFit	-	-	1	0,82
AI ResJobFit	+	-	2,3	0,81
AI ResJobFit	-	+	0,75	0,84
AI ResJobFit	+	+	1,72	0,83

Як видно з таблиці 5.5, використання автоматично згенерованої анотації в якості додаткової вхідної секції у модулі векторизації покращує показники якості без значного впливу на швидкість.

Висновки за розділом 5

1. Етап сегментації та виділення ключової інформації за запропонованою технологією дозволяє підвищити якість в порівнянні з базовою моделлю. Значення показника якості є вищим в порівнянні з

моделлю з комп'ютерним зором, а її швидкість при використанні методики дистиляції майже в 4 рази вище.

2. Нормалізація сутностей зазнала значного пришвидшення (в 13,4 рази в порівнянні з базовою технологією, яка використовує контекст для моделювання подань фраз, при цьому було досягнуто покращення якості на 20-50%)

3. Встановлено доцільність застосування дистиляції кожного технологічного етапу інформаційної технології AI ResJobFit, а також використання автоматично згенерованої анотації в якості додаткової вхідної секції.

4. Збільшення швидкості аналізу резюме та зіставлення з вимогами вакансій на 0,16 секунди для резюме та 0,15 для вакансії при застосуванні інформаційної технології AI ResJobFit без сумаризації в умовах відсутності рекрутера порівняно з технологією з використанням сумаризації пояснюється відсутністю використання додаткової енкодер-декодер моделі для генерації анотації. Якість технології без сумаризації знижується несуттєво порівняно з технологією із сумаризацією (в умовах без рекрутера), що обумовлює доцільність застосування інформаційної технології рекомендацій резюме в умовах відсутності рекрутера у варіанті без сумаризації.

5. При використанні інформаційної технології AI ResJobFit для інтенсифікації процесу відбору резюме рекрутером, швидкість ознайомлення рекрутера з документом (автоматично згенерованою анотацією) підвищується в 3,2 разів в порівнянні з швидкістю ознайомлення з повним текстом резюме для контрольного варіанту (без застосування сумаризації), і становить 2,7 та 7,9 секунд на один документ відповідно. Це вказує на значну інтенсифікацію процеса рекрутинга і дозволяє рекомендувати технологію інтенсифікації процесу відбору та ранжування резюме рекрутером у варіанті технології з сумаризацією.

ВИСНОВКИ

В результаті виконаних теоретичних та експериментальних досліджень було вирішено актуальну науково-практичну задачу предметно-орієнтованого аналізу природномовних текстів для автоматизації аналізу та інтерпретації великих масивів даних, що містяться в резюме та описах вакансій. Для цього було розроблено інформаційну технологію, яка дозволила підвищити ефективність рекрутингу шляхом збільшення як точності зіставлення кваліфікацій кандидатів з вимогами вакансій при мінімізації людського фактору, так і релевантності рекомендацій. Основними результатами даного дисертаційного дослідження є наступні:

1. В результаті проведеного аналізу процесів управління та рекрутингу в домені управління персоналом, а також сучасних методів та засобів інтелектуальної обробки текстової інформації встановлено необхідність підвищення ефективності таких процесів шляхом розробки та впровадження інтелектуальних інформаційних технологій автоматизованого аналізу документів, таких як резюме та вакансії, з використанням методів опрацювання природної мови.

2. Удосконалено модель представлення токенів для візуального насичених документів шляхом безпосередньої інтеграції параметрів про стильові ознаки (без використання комп’ютерного зору). Встановлено, що застосування безпосередньої інтеграції параметрів про стильові ознаки (без використання комп’ютерного зору) дозволяє удосконалити модель представлення токенів для візуального насичених документів. Це дозволило досягти покращення якості класифікації токенів в завданні сегментації резюме на 3-6% F1 метрики у порівнянні з адаптованою до домену версією моделі BERT, навченою за допомогою MLM стратегії навчання. В порівнянні з моделлю, яка використовує комп’ютерний зір, розроблена модель дозволяє досягти 1,67 пришвидшення обробки документів.

3. Запропоновано новий метод тренування подань назв посад, що базується на використання фраз навичок, які зазначені в описі роботи. Представлений підхід до нормалізації назв посад дозволяє досягти 10–21,5% покращення в завданні нормалізації назв посад порівняно з базовим (JobBERT). Створений еталонний набір даних щодо нормалізації назв посад був викладений в публічний доступ на Гітхабі.

4. Запропоновано структурну модель представлення фраз у контексті для їх групування та нормалізації, що базується на використанні спеціальних маркерів для виділення фраз, які моделюються. Новий метод тренування моделей з контекстно-орієнтованим вирівнюванням фраз призводить до покращення якості подання фраз на 9,9-10,6%. Встановлено, що Phrase2Vec безперебійно працює з новими навичками, які не були присутні в тренувальному наборі даних. Новий еталон семантичної подібності для навичок, витягнутих з реальних текстів з домену управління персоналом, разом з відповідними контекстними абзацами дозволяє надійно порівнювати моделі подання фраз навичок.

Використання розробленого методу подання фраз для їх нормалізації досягти пришвидшення в 7 разів в порівнянні з базовою моделлю та підвищення якості на 3-14%). За базову модель було використано McPhraSy, яка використовує контекст для подання фраз.

Створений еталонний набір даних був викладений в публічний доступ на Гітхабі.

Розроблений метод подання фраз у контексті було апробовано в компанії Data Science UA.

5. Розроблено метод некерованого навчання моделі HR Embedder з використанням структури документів щодо вивчення подань текстів у сфері управління персоналом. Метод базується на внутрішньо- та міжсекційному вирівнюванні подань текстів, що призвело до покращення NDCG на 11% порівняно зі стратегією DeCLUTR для завдання підбору резюме щодо

вакансій на основі автоматично згенерованих анотацій. Створений еталонний набір даних був викладений в публічний доступ на Гітхабі.

6. Розроблено метод векторизації структурованих документів в домені управління персоналом, який базується на синтетично згенерованих парах вакансія-резюме. Генерація вакансій проводиться з використанням BMM на основі останнього опису роботи, витягнутого з резюме. Цей метод дозволив досягти підвищення метрик якості на 2% в порівнянні з базовою моделлю (ConFit).

7. Розроблено метод зменшення обсягу тексту з урахуванням структури документа та ключових фраз для подальшої сумаризації. Встановлено доцільність скорочення тексту секцій проходячи речення за реченням та зберігаючи лише ключові фрази до досягнення бажаної довжини. Це дозволяє досягти 5% покращення якості сумаризації для довгих документів в домені управління людськими ресурсами в порівнянні з традиційною методикою, яка полягає в видалення стоп слів та найбільш частотних слів.

8. Розроблено метод некерованого попереднього тренування та застосування зваженої функції втрат для сумаризації документів з використанням їх структури. Використання структури документів та зваженої функції втрат з використання коефіцієнтів вагомості для токенів, які формують ключові фрази, дозволяє досягти 7% покращення при подальшому тренуванні на цільових даних. Особливо корисним є така стратегія попереднього тренування, коли кількість тренувальних зразків обмежена.

Встановлено поріг у 3 000 зразків для сумаризації резюме за допомогою згенерованих GPT-4 анотацій, після якого додаткові навчальні дані тієї ж якості не призводять до покращення продуктивності, і доцільно обмежити розмір навчального набору даних 3 000 зразками.

9. Отримала подальший розвиток інформаційна теорія для багатомовної дистиляції текстових подань на прикладів текстів українською мовою. Встановлено, що використання функції втрат на основі косинусної подібності призводить до значних покращень (на 14,2% за абсолютним показником NMI)

у порівнянні з використанням функції втрат середньоквадратичної похибки при дистиляції векторних подань текстів з добре натренованої моделі.

Встановлено, що вибір моделі-вчителя та розподіл косинусної подібності, який вона генерує, впливають на якість дистиляції векторних подань моделі-студента. Встановлена сильна негативна кореляція між “90-м перцентилем розподілу косинусних коефіцієнтів подібності” та “середнім показником NMI”, отриманим на еталонному наборі.

Створений еталонний набір даних для тестування текстових подань для української мови та викладений в публічний доступ на Гітхабі.

10. Розроблено засновану на методах глибоко навчання та архітектурах типу “Трансформер” інформаційну технологію обробки документів, яка дозволяє поєднати етапи сегментації, парсингу, векторизації в єдину інформаційну технологію для аналізу резюме та зіставлення з вимогами вакансій в умовах відсутності рекрутера. Параметри технології: швидкість обробки одного резюме складає 0,31 с, вакансії — 0,22; якість — 0,845 NDCG. Збільшення швидкості технології у варіанті без сумаризації пояснюється відсутністю необхідності використовувати додаткову енкодер-декодер модель для генерації анотації. Якість знижується несуттєво з 0,86 NDCG до 0,845 NDCG (порівняно з технологією у варіанті з сумаризацією).

11. Розроблено інформаційну технологію інтенсифікації процесу відбору та ранжування резюме рекрутером. Встановлено, що введення етапу сумаризації у технологічну схему предметно-орієнтованого аналізу дозволяє підвищити якість з 0,845 до 0,860 NDCG та прискорити процес ознайомлення рекрутера з резюме у 3,2 рази: час, який витрачається рекрутером на ознайомлення з резюме знижується з 7,9 секунд (без сумаризації) до 2,7 секунд (варіант технології з сумаризацією) за рахунок включення зручної для сприйняття анотації. Швидкість технології “AI ResJobFit” складає 0,45 секунди для резюме та 0,37 для вакансії при використанні графічної карти T4

та опрацюванні документів з використанням батчингу; якість встановлено на рівні 0,86 NDCG;

Розроблена інформаційна технологія використання штучних нейронних мереж для інтенсифікації процесу відбору та ранжування резюме рекрутером була апробована в компанії Daxtra Technologies.

12. Вдосконалено інформаційну технологію предметно-орієнтованого аналізу текстів шляхом застосування методу дистиляції моделей на кожному її етапі, що дозволило підвищити швидкість обробки документів (у 2,3 рази порівняно з не дистильованими моделями).

ПЕРЕЛІК ПОСИЛАНЬ

1. Bocharova, M. Y., & Malakhov, E. V. (2024, February). CapStyleBERT: Incorporating capitalization and style information into BERT for enhanced resumes parsing. In Proceedings of the 2024 13th International Conference on Software and Computer Applications (pp. 249-254). DOI: https://doi.org/10.1145/3651781.3651820 (Scopus)

2. Bocharova M.Y., Malakhov E.V. and Mezhuyev V.I., VacancySBERT: the approach for representation of titles and skills for semantic similarity search in the recruitment domain. Applied Aspects of Information Technology. 6, 1 (Apr. 2023), 52–59. DOI: https://doi.org/10.15276/aait.06.2023.4 ISSN: 2617-4316

3. Bocharova, M., Malakhov, E. (2024). Improving information theory of context-aware phrase embeddings in HR domain. Eastern-European Journal of Enterprise Technologies, 5 (2 (131)), 53–60. https://doi.org/10.15587/1729-4061.2024.313970 (Scopus) Q3

4. Bocharova M., Malakhov E., General-purpose text embeddings learning for Ukrainian language. Advanced Information Technology, vol.1(3), pp. 6–12, 2024. DOI: https://doi.org/10.17721/AIT.2024.1.01 ISSN: 2788-6603

5. Bocharova M.Y., Malakhov E.V., ResJobFit - end-to-end artificial neural networks based technology for job-resume matching. Applied Aspects of Information Technology. 2024; Vol. 7, No. 4: 378–391. DOI: https://doi.org/10.15276/aait.07.2024.27 ISSN: 2617-4316

6. Bocharova, M., & Malakhov, E. (2022). Analysis of neural techniques for learning sentence representations. In Інформатика, інформаційні системи та технології: тези доповідей XIX Всеукр. конференції студентів і молодих науковців (pp. 114–116). Одеса, Україна.

7. Bocharova, M. (2022). Trends and limitations in sentence embeddings learning. In 1st Student Scientific Conference of Joint Research Cooperation between Odesa I.I. Mechnikov National University and Huaiyin Institute of Technology (Online Edition), April 28–29, May 16, 2022.

8. Bocharova, M., & Malakhov, E. (2024). *Information technology in the AI-driven recruitment*. In *Тези доповідей, 76. Міжнародна науково-практична конференція "Сучасні інформаційні системи та технології в цифровому суспільстві"* (April 18–19, 2024). Харків, Україна.

9. Qin, C., Zhang, L., Cheng, Y., Zha, R., Shen, D., Zhang, Q., ... & Xiong, H. (2023). A comprehensive survey of artificial intelligence techniques for talent analytics. https://doi.org/10.48550/arXiv.2307.03195.

10. Abhishek Kaushik. Problems with Resume Screening - What You Should Do Instead? URL: https://www.wecreateproblems.com/blog/resume-screening (дата звернення: 25.02.2025).

11. Tiwari, A. & Nalamwar, S, "A review of resume analysis and job description matching using machine learning". International Journal on Recent and Innovation Trends in Computing and Communication. 2024; 12 (2): p. 247–250.

12. AI Recruitment Market: Global Industry Analysis and Forecast (2024-2030) URL: https://www.maximizemarketresearch.com/market-report/global-ai-recruitment-market/63261/ (дата звернення: 25.02.2025)

13. Vysotska, V., Nagachevska, O., Mozol, N., Chyrun, S., Chyrun, L., & Prokipchuk, O. (2024, October). Identifying Accent and Origin of English Language Users by Voice and Text Analysis, NLP and Machine Learning Technology. In 2024 IEEE 17th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET) (pp. 1-6). IEEE.

14. Andronati, O., Antoshchuk, S., Babilunha, O., Arsirii, O., Nikolenko, A., & Mikhalev, K. (2024, September). A method of constructing ensemble classifiers for recognizing audio data of various nature. In 2024 14th International Conference on Advanced Computer Information Technologies (ACIT) (pp. 758-761). IEEE. DOI: https://doi.org/10.1109/ACIT62333.2024.10712469

15. Andronati, O., Antoshchuk, S., Nikolenko, A., Arsirii, O., Babilunha, O., & Petrosiuk, D. (2023, September). Ensemble classifiers of audio data for speech

emotions recognition. In 2023 13th International Conference on Advanced Computer Information Technologies (ACIT) (pp. 623-626). IEEE. DOI: https://doi.org/10.1109/ACIT58437.2023.10275604

16. Концевой, А. О., & Бісікало, О. В. (2022). Моделі глибокого навчання для вирішення задач класифікації текстової інформації. Інформаційні технології та комп'ютерна інженерія.№ 3: 13–20.

17. Holubinka, D., Vysotska, V., Vladov, S., Ushenko, Y., Talakh, M., & Tomka, Y. (2025). Intelligent System for Recognizing Tone and Categorizing Text in Media News at an Electronic Business Based on Sentiment and Sarcasm Analysis. International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.17, No.1, pp. 90-139. DOI: https://doi.org/10.5815/ijieeb.2025.01.06

18. Кудрик, О. В., Бісікало, О. В., & Здітовецький, Ю. С. (2024). Методи тонкого налаштування штучного інтелекту. Вісник Вінницького політехнічного інституту, (4), pp. 139-146. DOI: https://doi.org/10.31649/1997-9266-2024-175-4-139-146

19. Jurafsky D., Martin J. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released January 12, 2025. https://web.stanford.edu/~jurafsky/slp3.

20. Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. Computational linguistics, 18(4), pp. 467-480.

21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, pp. 5998-6008.

22. Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and

Short Papers), (pp. 4171–4186), Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1423

23. Nguyen, T.B., Nguyen, Q.M., Nguyen, T.T.H., Do, Q.T., Luong, C.M. (2020) Improving Vietnamese Named Entity Recognition from Speech Using Word Capitalization and Punctuation Recovery Models. Proc. Interspeech 2020, 4263-4267. DOI: https://doi.org/10.21437/Interspeech.2020-1896

24. Huang, Y., Lv, T., Cui, L., Lu, Y., & Wei, F. (2022, October). Layoutlmv3: Pre-training for document ai with unified text and image masking. In Proceedings of the 30th ACM International Conference on Multimedia (pp. 4083-4091). DOI: https://doi.org/10.48550/arXiv.2204.083

25. Davis, B., Morse, B., Price, B., Tensmeyer, C., Wigington, C., & Morariu, V. (2022, October). End-to-end document recognition and understanding with dessurt. In European Conference on Computer Vision (pp. 280-296). Cham: Springer Nature Switzerland. DOI: https://doi.org/10.48550/arXiv.2203.166

26. Powalski, R., Borchmann, Ł., Jurkiewicz, D., Dwojak, T., Pietruszka, M., & Pałka, G. (2021). Going full-tilt boogie on document understanding with text-image-layout transformer. In Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16 (pp. 732-747). Springer International Publishing. DOI: https://doi.org/10.48550/arXiv.2102.09550

27. Garncarek, Ł., Powalski, R., Stanisławek, T., Topolski, B., Halama, P., Turski, M., & Graliński, F. (2021, September). Lambert: Layout-aware language modeling for information extraction. In International Conference on Document Analysis and Recognition (pp. 532-547). Cham: Springer International Publishing. DOI: 10.1007/978-3-030-86549-8_34

28. Bisikalo, O., Kovtun, O., & Kovtun, V. (2023, September). Neural network concept of ukrainian-language text embedding. In 2023 13th International Conference on Advanced Computer Information Technologies (ACIT) (pp. 566-569). IEEE. DOI: https://doi.org/10.1109/ACIT58437.2023.10275511

29. Войцеховський, В. В., & Бісікало, О. В. (2023). Методи та підходи у розробці технології індексації документів та їх застосування в cloud технологіях (Doctoral dissertation, ВНТУ).

30. Bojanowski, P., Grave E., Joulin, A. (2016). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, vol. 5. DOI: https://doi.org/10.1162/tacl_a_00051.

31. Reimers, N. & Gurevych, I. (2019). SentenceBERT: Sentence embeddings using siamese BERT networks. Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: China. (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410.

32. Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Kurzweil, R. (2018). Universal sentence encoder. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, (pp. 169–174), Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-2029

33. Carlsson, F., Gogoulou, E., Ylipää, E., Cuba Gyllensten, A., & Sahlgren, M. (2021). Semantic re-tuning with contrastive tension. In International Conference on Learning Representations, 2021.

34. Gao, T., Yao, X., & Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 6894–6910), Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.552

35. Wang, K., Reimers, N., & Gurevych, I. (2021). TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In Findings of the Association for Computational Linguistics: EMNLP (pp. 671–688), Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.59

36. Giorgi, J., Nitski, O., Wang, B., & Bader, G. (2020). Declutr: Deep contrastive learning for unsupervised textual representations. In Proceedings of the

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 879–895). Online. Association for Computational Linguistics. 10.18653/v1/2021.acl-long.72

37. Wang, S., Thompson L., Iyyer M. (2021). Phrase-BERT: Improved Phrase Embeddings from BERT with an Application to Corpus Exploration. Processing of conference on Empirical Methods in Natural Language Processing, 10837–10851. https://doi.org/10.18653/v1/2021.emnlp-main.846.

38. Cohen A., Gonen H., Shapira O., Levy R. (2022). McPhraSy: Multi-context phrase similarity and clustering. Findings of the Association for Computational Linguistics: EMNLP, 3538–3550. https://doi.org/10.18653/v1/2022.findings-emnlp.259

39. Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. Transactions of the association for computational linguistics, 8, 64-77.

40. Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

41. Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122. Association for Computational Linguistics.

42. Ganitkevitch J., Durme B., Callison-Burch C. (2013). PPDB: The Paraphrase Database. Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 758–764.

43. Liu S., Wang Z., Gao Y., Ren L. (2021). Human-centric relation segmentation: Dataset and solution. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 9, 4987-5001.

44. Aslanyan G., Wetherbee I., Aslanyan G. (2022). Patents phrase to phrase semantic matching dataset. PatentSemTech Workshop. https://doi.org/10.48550/arXiv.2208.01171

45. Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) (pp. 923–929)

46. Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, (pp. 1351–1361). Association for Computational Linguistics

47. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2020). Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), (pp. 878–891). Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.62

48. Wang, L., & Wei, F. (2022). Text embeddings by weakly-supervised contrastive pre-training. DOI: https://doi.org/10.48550/arXiv.2212.0353

49. Reimers, N., & Gurevych, I. (2020). Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation/ In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), (pp. 4512–4525), Online. Association for Computational Linguistics. 2020. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.365

50. Heffernan, K., & Schwenk, H. (2022). Bitext mining using distilled sentence representations for low-resource languages. Findings of the Association for Computational Linguistics: EMNLP, (pp. 2101–2112), Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2022.findings-emnlp.154

51. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

52. Lewis, M. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.703

53. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67. [T5]

54. Sammet, J., & Krestel, R. (2023, September). Domain-Specific Keyword Extraction using BERT. In Proceedings of the 4th Conference on Language, Data and Knowledge (pp. 659-665).

55. Song, M.; Feng, Y.; and Jing, L. 2023.A survey on recent advances in keyphrase extraction from pre-trained language models.Findings of the Association for Computational Linguistics: EACL 2023, 2153–2164.

56. Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International conference on machine learning (pp. 11328-11339). PMLR.

57. Wang, W., Bao, H., Huang, S., Dong, L., & Wei, F. (2021). Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics DOI: https://doi.org/10.18653/v1/2021.findings-acl.188

58. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. DOI: https://doi.org/arXiv:1910.01108.

59. Zu, S., & Wang, X. (2019). Resume information extraction with a novel text block segmentation algorithm. Int J Nat Lang Comput, 8, 29-48. DOI: https://doi.org/10.5121/ijnlc.2019.8503

60. Barducci, A., Iannaccone, S., La Gatta, V., Moscato, V., Sperlì, G., & Zavota, S. (2022). An end-to-end framework for information extraction from italian resumes. Expert Systems with Applications, 210, 118487. DOI: https://doi.org/10.1016/j.eswa.2022.118487

61. Sajid, H., Kanwal, J., Bhatti, S. U. R., Qureshi, S. A., Basharat, A., Hussain, S., & Khan, K. U. (2022, January). Resume parsing framework for e-recruitment. In 2022 16th International Conference on Ubiquitous Information Management and Communication (IMCOM) (pp. 1-8). IEEE. DOI: https://doi.org/10.1109/IMCOM53663.2022.9721762

62. Retyk, F., Fabregat, H., Aizpuru, J., Taglio, M., & Zbib, R. (2023). Resume Parsing as Hierarchical Sequence Labeling: An Empirical Study. RecSys in HR'23: The 3rd Workshop on Recommender Systems for Human Resources, in conjunction with the 17th ACM Conference on Recommender Systems, September 18–22, 2023, Singapore, Singapore. DOI: https://doi.org/10.48550/arXiv.2309.070

63. Espinal, A., Haralambous, Y., Bedart, D., & Puentes, J. (2023, May). A Format-sensitive BERT-based Approach to Resume Segmentation. In 2023 33rd Conference of Open Innovations Association (FRUCT) (pp. 30-37). IEEE. DOI: https://doi.org/10.23919/FRUCT58615.2023.10143072

64. Neculoiu, P., Versteegh, M. & Rotaru, M. "Learning text similarity with siamese recurrent networks". Proceedings of the 1st Workshop on Representation Learning for NLP. Berlin: Germany. 2016. p. 148–157. DOI: https://doi.org/10.18653/v1/W16-1617.

65. Tran, H. T., Vo, H. H. P. & Luu, S.T. "Predicting job titles from job descriptions with multi-label text classification". 8th NAFOSTED Conference on Information and Computer Science (NICS). 2021. DOI: https://doi.org/10.48550/arXiv.2112.11052.

66. Decorte, J.-J., Van Hautte, J., Demeester, T., Develder, C. (2021). JobBERT: Understanding job titles through skill. International workshop on Fair, Effective And Sustainable Talent management using data science (FEAST) as part of ECML-PKDD 2021. https://doi.org/10.48550/arXiv.2109.09605

67. Zbib, R., Lacasa Alvarez, L., Retyk, F. et al. “Learning job titles similarity from noisy skill labels”. 2022. DOI: https://doi.org/10.48550/arXiv.2207.00494.

68. Decorte, J.-J., Van Hautte, J., Deleu, J., Develder, C., Demeester, T. (2022). Design of negative sampling strategies for distantly supervised skill extraction. 2nd Workshop on Recommender Systems for Human Resources (RecSys in HR 2022) as part of RecSys 2022. DOI: https://doi.org/10.48550/arXiv.2209.05987

69. Bhola, A., Halder, K., Prasad, A., Kan, M.-Y. (2020). Retrieving Skills from Job Descriptions: A Language Model Based Extreme Multi-label Classification Framework. Proceedings of the 28th International Conference on Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.coling-main.513

70. Djumalieva, J., Sleeman, C. (2018). An Open and Data-driven Taxonomy of Skills Extracted from Online Job Adverts. Developing Skills in a Changing World of Work, 425–454. DOI: https://doi.org/10.5771/9783957103154-425

71. Decorte, J.-J., Verlinden, S., Van Hautte, J., Deleu, J., Develder, C., Demeester, T. (2023). Extreme Multi-Label Skill Extraction Training using Large Language Models. International workshop on AI for Human Resources and Public Employment Services (AI4HR&PES) as part of ECML-PKDD. DOI: https://doi.org/10.48550/arXiv.2307.10778

72. Clavié, B., & Soulié, G. (2023). Large Language Models as Batteries-Included Zero-Shot ESCO Skills Matchers. DOI: https://doi.org/10.48550/arXiv.2307.035

73. Mercan, Ö. B., Cavsak, S. N., Deliahmetoglu, A., & Tanberk, S. (2023, October). Abstractive text summarization for resumes with cutting edge NLP transformers and LSTM. In 2023 Innovations in Intelligent Systems and Applications Conference (ASYU) (pp. 1-6). IEEE.

74. Li, C., Fisher, E., Thomas, R., Pittard, S., Hertzberg, V., & Choi, J. D. (2020). Competence-level prediction and resume & job description matching using

context-aware transformer models. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) , pages 8456–8466, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.emnlp-main.679

75. Lavi, D., Medentsiy, V., & Graus, D. (2021). consultantbert: Fine-tuned siamese sentence-bert for matching jobs and job seekers. DOI: https://doi.org/arXiv:2109.06501.

76. Bhatia, V., Rawat, P., Kumar, A. & Shah, R. R. “End-to-End resume parsing and finding candidates for a job description using BERT”. 2019. DOI: https://doi.org/10.48550/arXiv.1910.03089.

77. Luo, Y., Zhang, H., Wen, Y. & Zhang, X. “Resumegan: an optimized deep representation learning framework for talent-job fit via adversarial learning”. In Proceedings of the 28th ACM international conference on information and knowledge management. 2019. pp. 1101–1110. DOI: https://doi.org/10.1145/3357384.3357899.

78. Yu, X., Zhang, J. & Yu, Z. . “ConFit: Improving resume-job matching using data augmentation and contrastive learning”. Association for Computing Machinery. 2024. p. 601–611. DOI: https://doi.org/10.1145/3640457.3688108.

79. Liu, S., Cao, J., Yang, R., & Wen, Z. (2022). Key phrase aware transformer for abstractive summarization. Information Processing & Management, 59(3), 102913. DOI: https://doi.org/10.1016/j.ipm.2022.102913

80. Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In Lrec (pp. 2214-2218).

81. Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., ... & Farhadi, A. (2022). Matryoshka representation learning. Advances in Neural Information Processing Systems, 35, 30233-30249. isbn: 9781713871088

82. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. DOI: https://doi.org/arXiv:1907.11692

83. “The malaysian national employment portal for all job seekers and employers”. – Available from: https://candidates.myfuturejobs.gov.my/search-jobs. – (дата звернення: 08.03.2023).

84. “ESCO: European skills, competences, qualifications and occupations”. EC Directorate E. 2017. – URL: https://esco.ec.europa.eu/en (дата звернення: 25.02.2025)

85. Atapour-Abarghouei, A., Bonner, S., & McGough, A. S. (2021, December). Rank over class: The untapped potential of ranking in natural language processing. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 3950-3959). IEEE.

86. Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2021). Deduplicating training data makes language models better. In proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8424–8445). Association for Computational Linguistics DOI: 10.18653/v1/2022.acl-long.577

87. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). Retrieval augmented language model pre-training. In Proceedings of the 37th International Conference on Machine Learning (pp. 3929–3938). JMLR.org.

88. “Why do recruiters spend only 7.4 seconds on resumes?”. Ladders Staff October 7, 2020 – URL: https://www.theladders.com/career-advice/why-do-recruiters-spend-only-7-4-seconds-on-resumes (дата звернення: 25.02.2025)

ДОДАТОК А

Список публікацій здобувача за темою дисертації

Наукова публікація у виданнях, що входять до міжнародних наукометричних баз Scopus та Web of Science:

1. Bocharova, M., Malakhov, E. (2024). Improving information theory of context-aware phrase embeddings in HR domain. Eastern-European Journal of Enterprise Technologies, 5 (2 (131)), 53–60. https://doi.org/10.15587/1729-4061.2024.313970 (Scopus) Q3

https://journals.uran.ua/eejet/article/view/313970

(Особистий внесок здобувача: запропоновано удосконалення інформаційної теорії контекстно-освіченого подання фраз, яке базується на використанні спеціальних токенів-маркерів для позначення меж фраз, створено датасет для оцінювання, проведено експерименти та оцінено результати)

Наукові публікації у фахових виданнях України:

2. Bocharova M.Y., Malakhov E.V. and Mezhuyev V.I., VacancySBERT: the approach for representation of titles and skills for semantic similarity search in the recruitment domain. Applied Aspects of Information Technology. 6, 1 (Apr. 2023), 52–59. DOI: https://doi.org/10.15276/aait.06.2023.4 ISSN: 2617-4316

http://dspace.opu.ua/xmlui/handle/123456789/13461

(Особистий внесок здобувача: запропоновано підхід до тренування подань назв посад, зібрано експериментальні дані та проведена оцінка результатів)

3. Bocharova M., Malakhov E., General-purpose text embeddings learning for Ukrainian language. Advanced Information Technology, vol.1(3), pp. 6–12, 2024 DOI: https://doi.org/10.17721/AIT.2024.1.01 ISSN: 2788-6603

https://ait.knu.ua/general-purpose-text-embeddings-learning-for-ukrainian-language/

(Особистий внесок здобувача: запропоновано удосконалення інформаційної теорії багатомовної дистиляції текстових подань на прикладі текстів українською мовою, орієнтуючись на визначення оптимальної функції втрат. Представлено новий датасет для тестування моделей векторних подань для української мови)

4. Bocharova M.Y., Malakhov E.V., ResJobFit - end-to-end artificial neural networks based technology for job-resume matching. Applied Aspects of Information Technology. 2024; Vol. 7, No. 4: 378–391. DOI: https://doi.org/10.15276/aait.07.2024.27 ISSN: 2617-4316

https://aait.od.ua/index.php/journal/article/view/265/265

(Особистий внесок здобувача: запропоновано інформаційну технологію для зіставлення вакансій та резюме у векторному просторі)

Наукові праці, які засвідчують апробацію матеріалів дисертації:

5. Bocharova, M. Y., & Malakhov, E. V. (2024, February). CapStyleBERT: Incorporating capitalization and style information into BERT for enhanced resumes parsing. In Proceedings of the 2024 13th International Conference on Software and Computer Applications (pp. 249-254). DOI: https://doi.org/10.1145/3651781.3651820

https://dl.acm.org/doi/10.1145/3651781.3651820

(Особистий внесок здобувача: запропоновано метод безпосереднього включення інформації про стильові ознаки в архітектуру типу “Трансформер”, проведено огляд літератури та проаналізовано результати)

6. Bocharova, M., & Malakhov, E. (2022). Analysis of neural techniques for learning sentence representations. In Інформатика, інформаційні системи та технології: тези доповідей XIX Всеукр. конференції студентів і молодих науковців (pp. 114–116). Одеса, Україна.

(Особистий внесок здобувача: досліджено сучасний стан моделей щодо подань текстів у векторному просторі)

7. Bocharova, M. (2022). Trends and limitations in sentence embeddings learning. In 1st Student Scientific Conference of Joint Research Cooperation

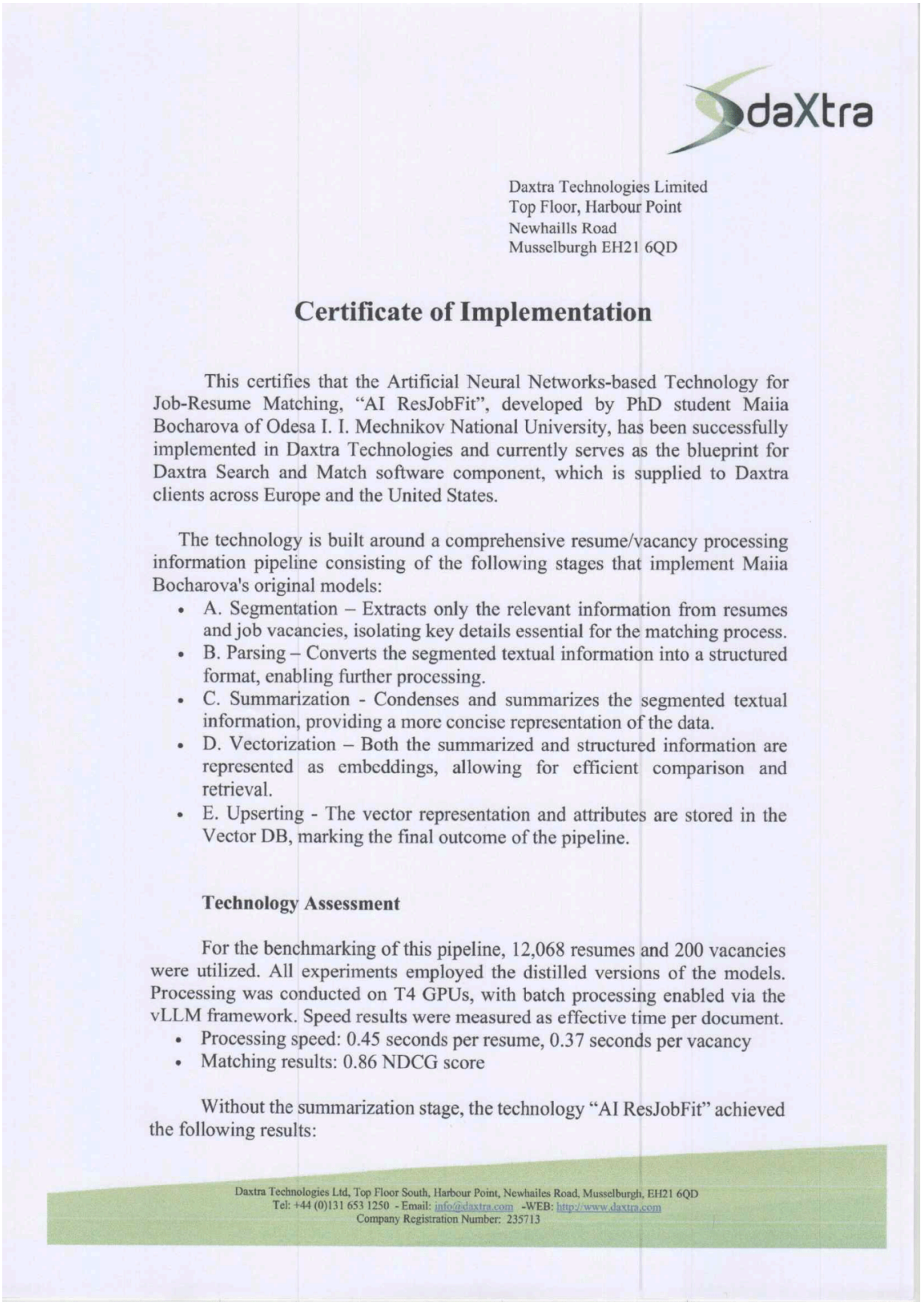
between Odesa I.I. Mechnikov National University and Huaiyin Institute of Technology (Online Edition), April 28–29, May 16, 2022.

(Особистий внесок здобувача: досліджено сучасний стан моделей щодо некерованого навчання подань текстів у векторному просторі)

8. Bocharova, M., & Malakhov, E. (2024). *Information technology in the AI-driven recruitment*. In *Тези доповідей, 76. Міжнародна науково-практична конференція "Сучасні інформаційні системи та технології в цифровому суспільстві"* (April 18–19, 2024). Харків, Україна.

(Особистий внесок здобувача: запропонована технологія виставлення вакансій та резюме та блоки, з яких ця технологія складається)

ДОДАТОК Б
Акт впровадження результатів у ТОВ “Daxtra Technologies”



Daxtra Technologies Limited
Top Floor, Harbour Point
Newhaills Road
Musselburgh EH21 6QD

Certificate of Implementation

This certifies that the Artificial Neural Networks-based Technology for Job-Resume Matching, “AI ResJobFit”, developed by PhD student Maiia Bocharova of Odesa I. I. Mechnikov National University, has been successfully implemented in Daxtra Technologies and currently serves as the blueprint for Daxtra Search and Match software component, which is supplied to Daxtra clients across Europe and the United States.

The technology is built around a comprehensive resume/vacancy processing information pipeline consisting of the following stages that implement Maiia Bocharova's original models:

- A. Segmentation – Extracts only the relevant information from resumes and job vacancies, isolating key details essential for the matching process.
- B. Parsing – Converts the segmented textual information into a structured format, enabling further processing.
- C. Summarization - Condenses and summarizes the segmented textual information, providing a more concise representation of the data.
- D. Vectorization – Both the summarized and structured information are represented as embeddings, allowing for efficient comparison and retrieval.
- E. Upserting - The vector representation and attributes are stored in the Vector DB, marking the final outcome of the pipeline.

Technology Assessment

For the benchmarking of this pipeline, 12,068 resumes and 200 vacancies were utilized. All experiments employed the distilled versions of the models. Processing was conducted on T4 GPUs, with batch processing enabled via the vLLM framework. Speed results were measured as effective time per document.

- Processing speed: 0.45 seconds per resume, 0.37 seconds per vacancy
- Matching results: 0.86 NDCG score

Without the summarization stage, the technology “AI ResJobFit” achieved the following results:

Daxtra Technologies Ltd, Top Floor South, Harbour Point, Newhailes Road, Musselburgh, EH21 6QD
Tel: +44 (0)131 653 1250 - Email: info@daxtra.com -WEB: http://www.daxtra.com
Company Registration Number: 235713

- Processing speed: 0.29 seconds per resume, 0.22 seconds per vacancy
- Matching results: 0.845 NDCG score

Component Assessment

The following advancements were noted in specific components of the technology:

- Segmentation and Key Information Extraction: Using the proposed style- and capitalization-aware model improved results by 3-6% over text-based models and by 1% with a 3x speedup over computer vision-based models. These results were validated on 1,500 resumes and 700 vacancies, manually annotated for sections and key entities.
- Context-Aware Phrase Representation Model: Achieved a 3-14% improvement in accuracy while maintaining a 7x speedup over the McPhraSy model. Validation was conducted on 12,000 human-annotated phrase pairs for classification and normalization tasks.
- Summarization: Using pretraining based on resume structure and key-phrase-aware loss, the technology achieved a 7% improvement in Rouge-1 score compared with the base BART model. This was validated on 700 human-checked resume summaries and 400 vacancy summaries.
- Vectorization Module: The vectorization approach resulted in a 5% improvement over the ConFit method. This was validated with 200 vacancies and 12,068 resumes, each marked in terms of relevance by recruiters.

Conclusion

The "AI ResJobFit" technology demonstrates superior performance in vacancy-resume matching compared to other AI-driven technologies evaluated at Daxtra, showcasing improved accuracy, speed, and overall effectiveness. As a result, this technology has been adopted as the blueprint for Daxtra's Search & Match software component.

28.01.2025

Andrei Mikheev, PhD,
Innovation Architect and Head of R&D Department
of Daxtra Technologies,

ДОДАТОК В

Акт апробації архітектури моделі векторного подання фраз для їх групування та нормалізації

Акт апробації
архітектури моделі векторного представлення фраз у контексті
для їх групування та нормалізації

Було досліджено архітектуру моделі векторного представлення фраз у контексті щодо її застосування для групування та нормалізації ключових фраз з нормативних документів в галузі будівництва.

Ця архітектура моделі базується на використанні специфічних токенів-маркерів ([PHRASE] та [/PHRASE] для позначення початку та кінця фрази відповідно.

При апробації архітектури моделі було використано нормативні документи “International Building Codes”.

Результати дозволяють підтвердити ефективність розробленої архітектури. Якість подання фраз із застосуванням нової архітектури було підвищено на 9.9% порівняно з класичною архітектурою.

20.01.2025

Директор Козирева В.В.

МП

ДОДАТОК Г
Акт про використання результатів в науково-дослідній роботі

«ЗАТВЕРДЖУЮ»
Проректор з наукової роботи
д-р біол. наук, професор
_____Володимир ІВАНИЦЯ
«___»_____ 2025 р.
мп

«ЗАТВЕРДЖУЮ»
Перший проректор
канд. політ. наук, професор
_____Майя НІКОЛАЄВА
«___»_____2025 р.
мп

АКТ
про використання результатів

дисертаційної роботи на здобуття наукового ступеня доктора філософії аспірантки кафедри математичного забезпечення комп'ютерних систем БОЧАРОВОЇ Майї Юріївни в науково-дослідній роботі, що виконується згідно тематичного плану Одеського національного університету імені І.І. Мечникова

Наукові та науково-практичні результати дисертаційної роботи БОЧАРОВОЇ Майї Юріївни використано в науково-дослідній роботі №311 «Методи, моделі, інформаційні технології розподілених систем підтримки прийняття організаційних рішень» (№ держреєстрації 0121U111663; науковий керівник НДР – д-р техн. наук, проф. Малахов Є.В.; відповідальний виконавець НДР – канд. техн. наук, доц. Пенко В.Г.).

Здобувачка брала участь у виконанні робіт за вказаною темою в якості виконавиці. В межах 4-го етапу (2024 р.) НДР здобувачкою розроблені: 1) архітектура контекстно-залежного моделювання подання фрази; 2) модель векторного представлення токенів для візуально насичених документів, в якій безпосередньо інтегруються параметри про стильові ознаки.

Запропонована модель CapStyleBERT показує високу продуктивність для всіх розмірів наборів даних, перевершуючи моделі, які не використовують інформацію про форматування тексту, на 3-6% для різної кількості навчальних зразків. Архітектура цієї моделі в поєднанні з новим підходом до навчання з контекстно-залежним вирівнюванням фраз призводить до значного покращення якості репрезентації фраз на 9,9-10,6 % залежно від конкретного сценарію.

Науковий керівник НДР №311
д-р техн. наук, проф. Євгеній МАЛАХОВ

Відповідальний виконавець НДР №311
канд. техн. наук, доц. Валерій ПЕНКО

ДОДАТОК Д

Акт про використання результатів в навчальному процесі кафедри

ЗАТВЕРДЖУЮ
Проректор з науково-педагогічної роботи
Одеського національного
університету імені І.І. Мечникова
канд. політ. наук, проф.

_____ Майя НІКОЛАЄВА
«___» _____ 2025 р.

АКТ

про використання результатів дисертаційної роботи на здобуття наукового ступеня доктора філософії аспірантки кафедри Математичного забезпечення комп'ютерних систем БОЧАРОВОЇ Майї Юріївни в навчальному процесі кафедри Одеського національного університету імені І.І. Мечникова.

Ми, що нижче підписалися, завідувач кафедри Математичного забезпечення комп'ютерних систем МАЛАХОВ Євгеній Валерійович та доцент кафедри Математичного забезпечення комп'ютерних систем ПЕНКО Валерій Георгійович склали акт про те, що результати дисертаційної роботи БОЧАРОВОЇ Майї Юріївни впроваджені у навчальний процес на кафедрі Математичного забезпечення комп'ютерних систем.

Архітектура контекстно-залежного подання фраз та алгоритм урізання тексту з урахуванням структури документа та ключових фраз для сумаризації впроваджено в дисципліні «Методи обробки текстів природної мови».

Завідувач кафедри МЗКС Євгеній МАЛАХОВ

Доцент кафедри МЗКС Валерій ПЕНКО