https://doi.org/10.31891/2307-5732-2026-363-60
УДК 004.8:004.912

НАТАЛІЧ ДЕНИС
Національний університет «Львівська політехніка»
https://orcid.org/0009-0005-8590-8302
e-mail: denys.natalich.mnsam.2024@lpnu.ua
ВАСИЛЮК АНДРІЙ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-3666-7232
e-mail: Andrii.S.Vasyliuk@lpnu.ua

ВИКОРИСТАННЯ МЕТОДІВ ОПРАЦЮВАННЯ ПРИРОДНОЇ МОВИ У ЧАТ-БОТАХ ДЛЯ ПОШУКУ ОБ'ЄКТІВ НЕРУХОМОСТІ

У роботі узагальнено результати проведеного комплексного аналізу ефективності застосування методів опрацювання природної мови у чат-ботах, призначених для автоматизації пошуку об'єктів нерухомості. Актуальність теми зумовлена швидким зростанням обсягів неструктурованих текстових даних, що надходять від користувачів у вигляді запитів та містяться в оголошеннях на різних онлайн-платформах.

Ключові слова: *чат-бот, нерухомість, опрацювання природної мови, семантичний пошук, трансформерні моделі.*

NATALICH DENYS, VASYLIUK ANDRII
Lviv Polytechnic National University

THE ROLE OF NLP IN CHATBOT SYSTEMS FOR REAL ESTATE BROKERAGE SUPPORT

This research investigates the use of modern natural language processing (NLP) models in conversational systems intended for automated real estate property retrieval. The motivation for this work stems from the increasing volume of unstructured textual information produced by users through natural-language queries, as well as descriptive content contained in online real estate listings. Conventional search systems based on exact keyword matching struggle to deliver relevant results, as they fail to account for semantic meaning, contextual interpretation, and linguistic diversity inherent in human communication.

To overcome these challenges, a chatbot-based prototype was developed, incorporating an NLP-driven query understanding module. The system is designed to infer user intent, identify key property-related attributes—such as property category, price constraints, location, number of rooms, and floor area—and apply semantic matching techniques to retrieve suitable listings.

An experimental evaluation was conducted using a custom-constructed dataset consisting of 520 real estate advertisements and 75 natural-language queries with varying levels of syntactic and semantic complexity. Three text representation and retrieval strategies were examined: a classical TF-IDF statistical model, a neural embedding approach using Word2Vec, and a contextual transformer-based model built on BERT.

The comparative analysis demonstrates that transformer-based representations provide a substantial advantage in semantic search tasks. The BERT-based approach achieved a Precision@5 score of 0.83, outperforming TF-IDF by 39% and Word2Vec by 24%.

The results show that the proposed chatbot system is capable of producing highly relevant property recommendations, reducing irrelevant matches, and improving the efficiency of natural-language query processing. These findings confirm the effectiveness of integrating state-of-the-art NLP techniques into real estate support systems and highlight promising avenues for future research, including the advancement of intelligent conversational agents and the expansion of semantic search technologies within the real estate domain.

Keywords: *chatbot, real estate, natural language processing, semantic search, transformer models.*

Стаття надійшла до редакції / Received 17.02.2026
Прийнята до друку / Accepted 03.03.2026
Опубліковано / Published 26.03.2026

Постановка проблеми

Сучасний ринок нерухомості характеризується значною динамікою розвитку, високою конкуренцією та постійним збільшенням обсягів інформації, що циркулює між учасниками ринку. У цій сфері щоденно генеруються тисячі нових оголошень про продаж і оренду об'єктів нерухомості, які містять значну кількість неструктурованої текстової інформації, зокрема, детальні описи житлових та комерційних приміщень, умови співпраці, технічні характеристики, особливості розташування, додаткові зручності тощо. До цього додаються повідомлення від власників, короткі та часто неточні запити від потенційних клієнтів, уточнення щодо деталей об'єктів, а також супровідне листування між брокером і користувачем. Усе це формує великий і різнорідний масив даних, який потребує швидкого й точного опрацювання.

Брокер або агент з нерухомості працює в умовах обмеженого часу і високої інформаційної насиченості. Він змушений переглядати значну кількість оголошень, порівнювати їх між собою, зіставляти з численними вимогами клієнтів та підтримувати ефективну комунікацію. Природні мовні формулювання клієнтів часто є неповними, багатозначними, емоційними або сформульованими у вільній манері. Замість конкретного структурованого запиту на кшталт «двокімнатна квартира у районі Сихів до 600 доларів», клієнт може написати: «Хочу щось недороге, ближче до центру, але не обов'язково, головне — дві окремі кімнати». Така форма подання інформації є природною для людини, але вкрай складною для автоматизованого опрацювання.

Традиційні інформаційно-пошукові системи, що працюють на основі порівняння ключових слів, виявляються недостатньо ефективними у сфері нерухомості. Вони не враховують семантичні зв'язки між словами, не розпізнають контекст використання термінів, ігнорують синонімію та стилістичні варіації, а також не

здатні зрозуміти приховані вимоги. Наприклад, система може трактувати запит «квартира біля вокзалу» як будь-який об'єкт, у тексті якого згадується слово «вокзал», навіть якщо опис стосується іншого району або використовується лише умовний вираз. Такий підхід призводить до формування надто великих та нерелевантних добірок, що ускладнює роботу брокера, збільшує час пошуку об'єктів і знижує якість взаємодії з клієнтом.

На цьому тлі чат-боти, інтегровані в популярні месенджери, стають дедалі більш поширеним інструментом у сфері обслуговування клієнтів. Вони здатні забезпечити цілодобову взаємодію, оперативно реагувати на запити та автоматизувати рутинні комунікаційні процеси. Проте більшість існуючих чат-ботів у сфері нерухомості обмежуються простими сценаріями, такими як пересилання інформації, збирання контактних даних або застосування примітивних фільтрів. Такі системи не здатні інтерпретувати повноцінні природномовні запити і не можуть забезпечити семантичний аналіз тексту, унаслідок чого їх практичне використання стає менш ефективним.

Внаслідок цього, виникає наукова та практична проблема, яка полягає у необхідності створення інтелектуального чат-бота, здатного автоматично опрацьовувати природномовні запити користувачів, розпізнавати ключові параметри нерухомості, виділяти приховані смислові зв'язки та здійснювати семантичний пошук у базах даних об'єктів. Для цього потрібно впровадити сучасні методи опрацювання природної мови, оцінити їх ефективність на реальних даних, порівняти точність різних моделей та визначити оптимальний підхід для автоматизованого пошуку.

Об'єктивна потреба у підвищенні точності, релевантності та швидкодії пошукових систем у сфері нерухомості зумовлює актуальність цього дослідження. Практичне впровадження NLP-модулів у чат-боти може суттєво покращити процес підбору нерухомості, зменшити навантаження на брокерів і підвищити якість сервісу. Саме це обґрунтовує необхідність науково обґрунтованої розробки та експериментальної перевірки NLP-рішень для автоматизації задачі пошуку нерухомості.

Аналіз досліджень та публікацій

Тема застосування чат-ботів у бізнес-процесах активно досліджується впродовж останнього десятиліття. У працях зарубіжних та вітчизняних авторів відзначається, що чат-боти дають змогу автоматизувати типові сценарії взаємодії з клієнтами, зменшити навантаження на операторів та підвищити швидкість опрацювання звернень [1, 2]. Показано, що впровадження чат-ботів як першої лінії підтримки дозволяє скоротити до 30–40 % рутинних запитів, забезпечуючи цілодобову доступність сервісу та стабільну якість відповідей.

У дослідженнях [2] запропоновано підходи до оцінювання якості інтелектуальних діалогових агентів, зокрема за критеріями коректності відповідей, зрозумілості діалогу та задоволеності користувачів. Окрім цього, в роботах [4] розглядають архітектури сучасних чат-ботів, акцентуючи увагу на поєднанні модулів керування діалогом із методами штучного інтелекту, зокрема алгоритмами опрацювання природної мови.

Окремий напрям становлять дослідження в галузі методів опрацювання природної мови (ОПМ). Класичні підходи, такі як TF-IDF та модель «мішка слів» (Bag-of-Words), широко застосовуються для базового аналізу текстів і побудови інформаційно-пошукових систем. Їхнім недоліком є відсутність урахування семантичних зв'язків між словами та контексту. Розвиток розподілених векторних представлень слів (Word2Vec, GloVe) [5] став важливим кроком уперед, оскільки дозволив моделювати семантичну близькість між термінами та покращити якість пошуку й класифікації текстів.

Подальший якісний стрибок пов'язаний із появою трансформерних моделей, зокрема BERT [6], які забезпечують контекстуальне представлення слів та фраз, враховуючи їхнє оточення в реченні. Такі моделі продемонстрували високу точність у задачах семантичного пошуку, класифікації текстів та відповіді на запитання, що зумовило їх активне впровадження у діалогові системи нового покоління.

Застосування штучного інтелекту в сфері нерухомості розглядається, зокрема, у роботах [7], де аналізуються можливості використання аналітичних моделей для прогнозування цін та оцінювання інвестиційної привабливості об'єктів. Однак більшість цих досліджень фокусується на аналітиці структурованих даних (цінові ряди, характеристики об'єктів) і значно менше уваги приділяє неструктурованим текстам оголошень та природномовним запитам користувачів.

Аналіз світового досвіду свідчить про наявність прикладів використання інтелектуальних систем у домені нерухомості (наприклад, комерційні платформи з рекомендаційними сервісами), проте їхні алгоритми закриті, а результати досліджень — фрагментарні. При цьому питання порівняльної оцінки різних NLP-підходів саме в задачі семантичного пошуку нерухомості на основі природномовних запитів практично не висвітлено.

Проведений огляд наукових праць показує, що, попри значний прогрес у розвитку чат-ботів та методів опрацювання природної мови (ОПМ), комплексні експериментальні дослідження, спрямовані на поєднання цих технологій для автоматизації пошуку об'єктів нерухомості, залишаються недостатньо розробленими. Це обґрунтовує актуальність і новизну запропонованого у статті підходу.

Формулювання цілей статті

Враховуючи виявлені у процесі аналізу проблеми застосування традиційних підходів до пошуку нерухомості та обмеженість існуючих чат-ботів у сфері опрацювання природної мови, **метою роботи є** розроблення, реалізація та експериментальна перевірка ефективності NLP-модуля для чат-бота, який здатний інтерпретувати природномовні запити користувачів та здійснювати семантичний пошук об'єктів нерухомості. З метою досягнення поставленої цілі заплановано виконання таких взаємопов'язаних завдань: 1) сформувати

репрезентативний датасет оголошень про нерухомість та природномовних запитів користувачів, який може бути використаний для навчання та тестування NLP-модуля; 2) розробити прототип чат-бота, інтегрованого з модулем опрацювання природної мови, що здатний визначати ключові параметри нерухомості з текстових запитів, включаючи тип об'єкта, діапазон цін, кількість кімнат, місце розташування та додаткові характеристики; 3) реалізувати й порівняти три підходи до текстового пошуку — TF-IDF, Word2Vec та BERT — у контексті задачі зіставлення описів нерухомості з користувацькими запитами; 4) провести експериментальне оцінювання моделей, визначивши їхню точність, швидкодію та релевантність результатів за показниками Precision@5, MRR та середнім часом відповіді; 5) проаналізувати отримані результати та визначити модель, яка забезпечує найвищу ефективність у задачі пошуку нерухомості та доцільна для практичного впровадження.

Таким чином, ціль статті полягає не лише в теоретичному огляді можливостей NLP, а насамперед у практичній демонстрації й науково обґрунтованій оцінці ефективності конкретних алгоритмів у контексті автоматизації роботи брокера з нерухомості.

Виклад основного матеріалу

У процесі створення інтелектуального чат-бота для автоматизації пошуку об'єктів нерухомості було визначено, що найбільш критичним чинником є здатність системи працювати з природною мовою користувачів, яка у реальних умовах має довільний характер, включає неоднозначні формулювання, розмовні конструкції, емоційні вставки та неповні характеристики. Зазвичай люди, звертаючись до брокера, не описують свої потреби у формі системи фільтрів, як це відбувається на класичних сайтах з оренди чи продажу нерухомості. Навпаки, вони формулюють запит так, як їм зручно: «хочу щось ближче до центру, але не дуже дороге», «шукаю квартиру, щоб була тиха вулиця», «потрібне офісне приміщення біля вокзалу, але не в самій будівлі вокзалу», «квартира з двома спальнями, можна невелику, але світлу». Кожен із таких запитів є цілісною мовною конструкцією, яка містить змішану інформацію: частина є структурованими критеріями (дві спальні, близькість до вокзалу, обмеження за ціною), а частина — описовими побажаннями або емоційними компонентами, які традиційні пошукові системи не здатні коректно опрацьовувати.

Для того щоб чат-бот міг працювати з такими запитами, необхідно було створити відповідний текстовий корпус, який би охоплював не лише самі оголошення про нерухомість, але й реальні зразки мовлення користувачів. Корпус оголошень формувався на основі відкритих онлайн-платформ, що містять значний обсяг пропозицій у різних містах. Оголошення мали різнорідну структуру - від коротких "одним рядком" до довгих, детально оформлених текстів із великими описовими вставками. Частина оголошень включала структуровані дані, такі як площа, кількість кімнат, наявність балкону чи мансарди, однак у багатьох випадках інформація подавалася хаотично, без єдиного формату. Досить часто зустрічалося дублювання інформації або "розмиті" формулювання: «простора квартира», «гарний ремонт», «тихий район», що не дають чітких числових характеристик, але впливають на пошук.

Для забезпечення цілісності корпусу було проведено очищення текстів. Воно включало видалення HTML-тегів, рекламних вставок, емодзі, контактних даних, зайвих пробілів, повторів та інших шумових елементів. Після цього проводилася лематизація українських текстів, що дозволило зменшити кількість словоформ і зробити словникове представлення корпусу більш компактним. На цьому ж етапі формувався список стоп-слів, які не несуть значущого внеску у семантичний аналіз. Окремо були додані доменно-специфічні стоп-слова, що часто зустрічаються у оголошеннях ("зручний", "комфортний", "терміново", "актуально", "чудова пропозиція"), але не визначають характеристик об'єкта нерухомості.

Паралельно формувалися природні користувацькі запити. До цього процесу було залучено брокерів, які описали своє реальне спілкування з клієнтами з досвіду. У результаті було створено 75 варіантів запитів, які включали як прості («однокімнатна до 10 тис»), так і складні («квартира для молодої сім'ї, щоби поруч була школа, і бажано світла кухня»). Такі запити давали можливість перевірити, наскільки добре NLP-система справляється не лише з ключовими словами, а й з розширеним контекстом, натяками, прихованими умовами та неявними вимогами.

Наступним етапом стало розроблення архітектури чат-бота. Архітектуру було побудовано за модульним принципом, що дозволило окремо оптимізувати кожну частину системи. Комунікаційний модуль взаємодіяв з користувачем через Telegram API, розпізнавав тип повідомлення, виявляв помилки, формував відповіді або уточнювальні питання. Модуль попереднього опрацювання тексту відповідав за токенізацію, лематизацію, очищення, виділення числових та географічних сутностей. Особлива увага приділялася ідентифікації районів, які можуть бути позначені у різний спосіб: "Сихів", "Syhiv". Ще одним завданням було вилучення параметрів типу "не перший поверх", "неподалік парку", "без меблів", які вимагали коректного семантичного тлумачення.

Після етапу первинної підготовки даних та розроблення механізмів попереднього опрацювання тексту - основну увагу було зосереджено на створенні NLP-модуля, який відповідає за інтерпретацію природномовних запитів і знаходження релевантних об'єктів у базі оголошень. Саме цей модуль є "інтелектуальним ядром" системи, оскільки його ефективність визначає, наскільки чат-бот буде здатний виконувати пошук об'єктів так само точно, як це робить професійний брокер. Для того щоб модуль працював коректно, необхідно було не лише векторизувати тексти, але й навчитися виокремлювати всі явні та неявні ознаки запиту, зіставляти їх із

семантичними характеристиками оголошень і враховувати контекст, синонімію, стилістичні варіації й приховані смисли.

Крім цього, у межах дослідження, з метою детального відображення внутрішньої логіки роботи чат-бота, було побудовано діаграму послідовності рис. 1, яка демонструє взаємодію між усіма ключовими компонентами системи. На діаграмі показано повний цикл опрацювання природномовного запиту – від моменту його надсилання користувачем у Telegram до формування ранжованої добірки релевантних оголошень.

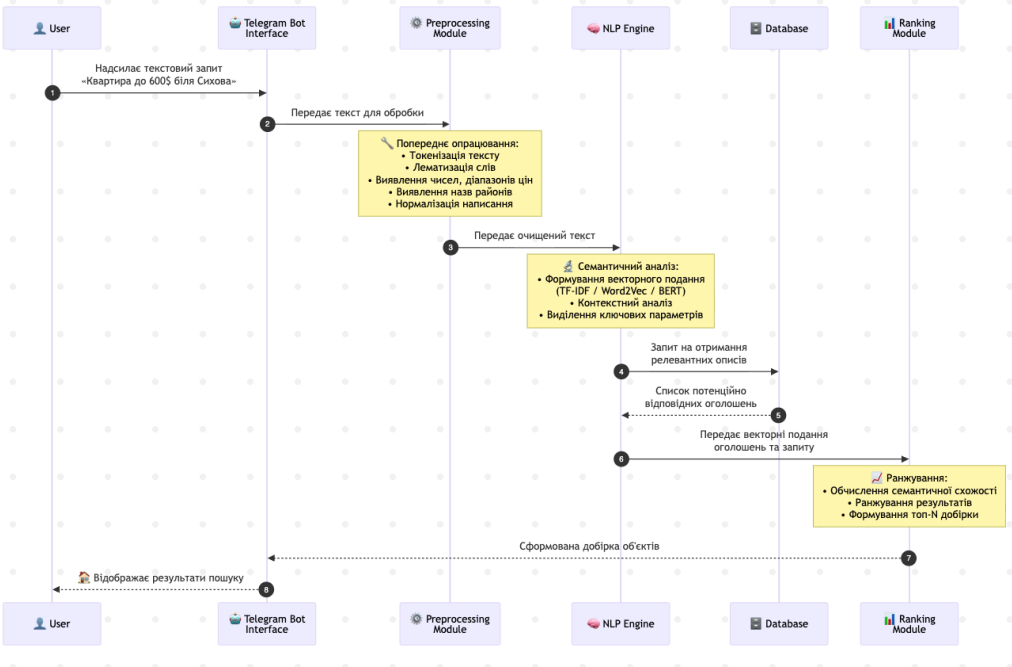

Рис. 1. Взаємодія між усіма ключовими компонентами системи

Після того як користувач надсилає текстове повідомлення, інтерфейс чат-бота отримує його та передає у модуль попереднього опрацювання тексту. На цьому етапі виконується токенізація, очищення, лематизація, нормалізація та виділення доменно-специфічних сутностей, зокрема цінових діапазонів, кількості кімнат, назв районів чи географічних орієнтирів. Отриманий уніфікований текст передається до NLP-модуля, який генерує векторне представлення запиту. Залежно від використовуваного підходу (Term Frequency – Inverse Document Frequency (TF-IDF), Word2Vec або BERT) виконується або статистичне оцінювання важливості слів, або побудова семантичних векторів, або ж повноцінне контекстуальне моделювання структури речення.

Після формування векторного подання системі необхідно зіставити його з описами об’єктів у базі даних. NLP-модуль надсилає запит до бази оголошень, отримує відповідні тексти, векторизує їх та передає в модуль ранжування. Саме там обчислюється міра семантичної подібності між запитом та описами, що дає змогу сформувати впорядкований список варіантів. Після цього добірка повертається у чат-бот, який відправляє результати користувачу.

Таким чином, діаграма послідовності відображає ключову особливість системи: взаємодія між її компонентами має строго послідовний характер, що забезпечує узгодженість усіх етапів — від попереднього опрацювання до формування релевантних рекомендацій. Саме така поетапна структура дозволяє досягти високої точності пошуку в умовах роботи з природномовними запитами.

У розробленій системі було реалізовано три різні підходи до побудови векторних представлень тексту TF-IDF, Word2Vec та BERT. Використання кількох моделей одночасно дозволило оцінити сильні та слабкі сторони кожної з них і визначити, яка з моделей найкраще відповідає специфіці ринку нерухомості, де запити часто містять неформальні або неповні описи. TF-IDF був обраний як базовий варіант через простоту реалізації та швидкість роботи. Він аналізував тексти на основі частоти зустрічання термінів, що дозволяло отримати первинне уявлення про схожість між запитом та оголошенням. Модель TF-IDF грунтується на ваговій мірі, яка оцінює важливість терміна t у документі d відносно всього корпусу документів D. Міра складається з двох компонентів TF — наскільки часто термін зустрічається в документі, IDF — наскільки рідкісний термін у всьому корпусі:

$$\text{TF-IDF}(t, d, D) = \frac{f_{t,d}}{|d|} \cdot \log \frac{N}{n_t}, \quad (1)$$

де $f_{t,d}$ — частота терміна у документі, $|d|$ — довжина документа, N — кількість документів у корпусі D, n_t — кількість документів, що містять термін t.

TF вимірює важливість терміну *всередині документа*. IDF вимірює важливість терміна *в межах всього корпусу*. Разом TF-IDF виділяє терміни, що часто зустрічаються у конкретному документі, але рідко в інших. Проте модель не враховує порядок слів і складну семантику, що часто призводить до поверхневого розуміння запиту. Наприклад, TF-IDF однаково опрацьовував фрази «квартира біля вокзалу» та «квартира з видом на вокзал», хоча їхній зміст суттєво різниться. Це зумовлювало низьку релевантність при роботі зі складними запитами.

Модель Word2Vec продемонструвала суттєво кращі результати завдяки здатності навчатися на великих корпусах текстів і формувати щільні векторні уявлення слів, у яких семантично подібні поняття розташовуються близько один до одного у просторі R^d. Формально Word2Vec реалізує одну з двох архітектур — Skip-Gram або CBOW. У моделі Skip-Gram параметри $\theta = \{v_w, u_w\}$ навчаються так, щоб максимізувати ймовірність появи контекстних слів w_{t+j} для заданого центрального слова w_t:

$$\max_{\theta} \sum_{t=1}^{T} \sum_{\substack{-c \le j \le c \\ j \neq 0}} \log p\left(w_{t+j} \mid w_t\right), \tag{2}$$

де умовна ймовірність визначається через softmax:

$$p(w_O \mid w_I) = \frac{\exp\left(u_{w_O}^{\top} v_{w_I}\right)}{\sum_{w=1}^{V} \exp\left(u_w^{\top} v_{w_I}\right)}, \tag{3}$$

У разі CBOW, навпаки, модель передбачає центральне слово за середнім вектором його контексту:

$$p(w_t \mid C_t) = \frac{\exp\left(u_{w_t}^{\top} v_{C_t}\right)}{\sum_{w=1}^{V} \exp\left(u_w^{\top} v_{C_t}\right)}, \tag{4}$$

$$v_{C_t} = \frac{1}{2c} \sum_{w \in C_t} v_w. \tag{5}$$

Оскільки обчислення softmax над усім словником є дорогим, на практиці застосовується **Negative Sampling**, що мінімізує втрати:

$$\mathcal{L}_{NS} = -\log \sigma\left(u_{w_O}^{\top} v_{w_I}\right) - \sum_{i=1}^{k} \log \sigma\left(-u_{w_i}^{\top} v_{w_I}\right), \tag{6}$$

де $\sigma(x) = \frac{1}{1+e^{-x}}$, а w_I — негативні приклади, вибрані з шумового розподілу.

Завдяки такій математичній структурі Word2Vec коректно опрацьовує синонімічні або морфологічно споріднені слова. Наприклад, векторне представлення слів «центр» та «центральний» розташовується на близькій відстані, що дозволяє моделі узагальнювати запити користувачів із різними формулюваннями.

Водночас ключовим недоліком Word2Vec є те, що модель працює на рівні окремих слів і не моделює фразову або реченнєву структуру. Тому контекст часто враховується лише частково. Наприклад, запит «квартира не в центрі, але недалеко від нього» містить заперечення та умовний зворот, які Word2Vec інтерпретує не як цілісну семантичну конструкцію, а як набір незалежних слів. Це може призвести до неправильного розуміння наміру користувача в складних ситуаціях.

Трансформерна модель BERT виявилася найефективнішою серед трьох підходів. Її архітектура базується на двонаправленому самоуваговому механізмі, який дозволяє моделі одночасно аналізувати речення зліва направо і справа наліво. Формально прихований стан *i*-го токена визначається як:

$$h_i = \text{TransformerLayer}(x_1, x_2, \ldots, x_n), \tag{7}$$

де $x_1, x_2, \ldots, x_n$ — вхідні токени, а кожен шар трансформера оновлює представлення слова з урахуванням усіх інших позицій.

Сам механізм уваги для токена і обчислюється за стандартною формулою self-attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{8}$$

де $Q = XW^Q, K = XW^K, V = XW^V$, а d_k – розмірність простору ключів. У випадку BERT використовується мультикомпонентна увага:

$$\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{H})W^O, \tag{9}$$

$$\text{head}_\text{j} = \text{Attention}\left(XW_j^Q, XW_j^K, XW_j^V\right). \tag{10}$$

Саме завдяки цьому BERT розглядає кожне слово в контексті всіх інших слів у реченні та виявляє складні семантичні залежності.

Під час тестування модель продемонструвала здатність коректно інтерпретувати навіть складні природномовні запити. Формально задачу класифікації або пошуку релевантних параметрів можна представити як:

$$y = \text{softmax}\left(W\, h_{[\text{CLS}]} + b\right), \tag{11}$$

де $h_{[\text{CLS}]}$ — контекстуальне представлення всього запиту, а y — прогнозований розподіл імовірностей.

Наприклад, запит *«квартира для сім'ї з дитиною в безпечному районі»* містить імпліцитні ознаки, які не подані явно. Якщо позначити узагальнений вектор прихованих потреб користувача як:

$$z = f\left(h_{[\text{CLS}]}\right), \tag{12}$$

то модель автоматично ідентифікує латентні характеристики, зокрема:

- необхідність додаткової кімнати:

$$P(\geq 2 \text{ кімнати} \mid z) = \sigma(w_1^{\top} z), \tag{13}$$

- вимогу до характеристик «безпечного району»:

$$P(\text{безпечний район} \mid z) = \sigma(w_2^{\top} z), \tag{14}$$

де σ — сигмоїда.

BERT також здатна пов'язувати описові фрази з узагальненими поняттями. Наприклад, множина контекстних фраз {"тихий район", "спокійна вулиця", "розвинена інфраструктура"} може бути віднесена моделлю до спільного латентного фактору безпека через високі увагові коефіцієнти:

$$\alpha_{i \to j} = \text{softmax}\left(\frac{q_i k_j^T}{\sqrt{d_k}}\right), \tag{15}$$

де великі значення $\alpha_{i \to j}$ вказують на семантичну пов'язаність токенів.

Таким чином, BERT не просто знаходить збіги слів, а будує контекстуальні семантичні представлення, які дозволяють коректно інтерпретувати натяки, приховані умови та описові елементи запитів.

У межах експерименту кожен запит було зіставлено з усією базою оголошень. Для порівняння результатів використовувалися стандартні метрики інформаційного пошуку: Precision@5 та Mean Reciprocal Rank (MRR). Precision@5 давав змогу оцінити, наскільки точними є перші п'ять результатів пошуку, оскільки саме вони мають найбільше значення для користувача. MRR визначав, на якій позиції в рейтингу з'являється перший релевантний результат, що дозволяло оцінити, наскільки добре модель ранжує знайдені об'єкти.

Додаткову увагу в дослідженні було приділено аналізу підходів до трансферного навчання в опрацюванні природної мови, які стали фундаментом ефективності моделей BERT та інших трансформерних архітектур. Важливим теоретичним орієнтиром слугував оглядовий матеріал аналітичної лабораторії Fast Forward Labs, де докладно описано еволюцію моделей у напрямі від базових статистичних методів до сучасних глибоких контекстно-орієнтованих представлень, а також показано характерні приклади впливу трансферного навчання на точність NLP-систем у різних доменах [8].

На рис. 2 наведено приклад, який демонструє, як попереднє навчання великих мовних моделей на масштабних корпусах даних значно підвищує ефективність роботи з вузькоспеціалізованими завданнями, навіть коли доступна навчальна вибірка є обмеженою. Цей принцип напряму корелює з результатами нашого дослідження: модель BERT, попередньо навчена на багатомовних корпусах, змогла забезпечити найвищу точність семантичного пошуку в задачах аналізу реальних текстів оголошень про нерухомість. Відповідну візуалізацію з роботи Fast Forward Labs доцільно використати для ілюстрації переваг трансферного навчання в межах запропонованої системи чат-бота.

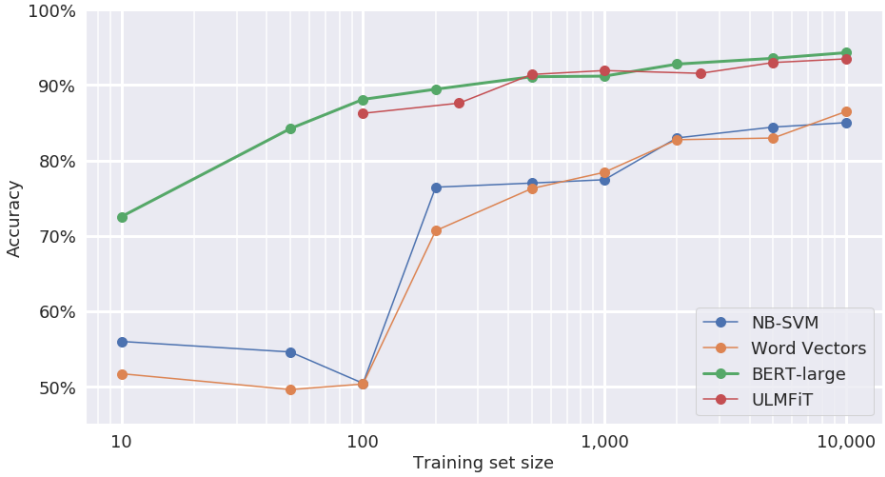

Рис. 2. Порівняння точності різних NLP-моделей залежно від розміру навчальної вибірки

Результати виявили суттєву різницю в ефективності моделей. TF-IDF показав Precision@5 приблизно 0,44, що означає невелику релевантність у більшості сценаріїв. Word2Vec досяг Precision@5 = 0,59, що вже можна вважати задовільним результатом для базового аналізу текстів. Однак саме BERT забезпечив Precision@5 = 0,83, що свідчить про значну точність семантичного пошуку. Значення MRR підтвердили цю тенденцію: для TF-IDF воно становило 0,38, для Word2Vec — 0,52, для BERT — 0,79. Таким чином, трансформерна модель дозволяла користувачу отримувати релевантні рекомендації у перших рядках списку, що суттєво економило час і покращувало взаємодію.

Таблиця 1

Порівняння ефективності NLP-моделей для семантичного пошуку нерухомості

Модель	Підхід до представлення тексту	Precision@5	MRR	Середній час відповіді (мс)	Стійкість до орфографічних помилок	Опрацювання складних/контекстних запитів
TF-IDF	Статистичний (частота термінів)	0.44	0.38	75	Низька	Низька
Word2Vec	Векторні представлення (CBOW / Skip-Gram)	0.59	0.52	93	Середня	Обмежена
BERT	Контекстуальні вектори (трансформер)	0.83	0.79	112	Висока	Висока

Особливо важливим моментом було тестування моделей на “нечітких” або “неповних” запитах. У реальному житті користувач може не знати точної кількості кімнат, не пам’ятати площу або сформулювати свої побажання у вигляді загального опису. Тестування показало, що TF-IDF зазвичай “губився” у таких випадках, Word2Vec справлявся частково, а BERT демонстрував стабільно високу ефективність.

Додатковим напрямом аналізу стало дослідження того, наскільки різні моделі здатні працювати із запитами, що містять у собі складні конструкції, подвійні умови, багатокомпонентні вимоги або контекстуально залежні характеристики. Наприклад, запит «квартира для сім’ї, але неподалік центру, щоб дитині було зручно добиратися до школи» включає кілька важливих параметрів, які не зазначені у структурованій формі. Слова “сім’я”, “зручно добиратися”, “школа” не є прямими ознаками у жодному оголошенні, але містять прихований зміст: зазвичай такі запити пов’язані з наявністю поблизу дитячих освітніх закладів, розвиненою інфраструктурою, доступністю громадського транспорту, а також із потребою в більшій кількості житлових кімнат. Трансформерні моделі, зокрема BERT, показали здатність виявляти ці приховані смисли. Вони враховують не лише окремі слова, але й їх взаємозв’язки в межах довших речень, що дозволяє формувати більш точні підбірки.

Під час експериментального тестування було перевірено також запити, які містять негативні формулювання та логічні виключення, з якими звичайні пошукові системи часто не справляються. Наприклад, фраза «не на першому поверсі, але без ліфта теж небажано» вимагає складної логічної інтерпретації. TF-IDF та Word2Vec зазвичай опрацьовували такий запит некоректно: TF-IDF звертав увагу лише на слова “перший”, “поверх”, “ліфт”, але не на зв’язки між ними, а Word2Vec ігнорував логічні оператори. Натомість модель BERT правильно розпізнавала смислові зв’язки між частинами запиту, відокремлювала “заборонені” умови від “бажаних” і в результаті пропонувала лише ті об’єкти, які відповідали обом критеріям одночасно.

Окрему увагу приділено тестуванню моделей на описах, що містять географічні особливості. В оголошеннях нерухомості часто використовуються назви районів, мікрорайонів, вулиць, торгівельних центрів, станцій метро або локальних орієнтирів. Фактором, що ускладнює процес є те, що користувачі можуть вводити назви у різних формах: «Сихів», «Syhiv», «на Сихові», «в районі Сихова». Система мала об’єднати ці варіанти в єдину сутність. Під час тестування TF-IDF реагував на такі відмінності як на різні слова, Word2Vec частково розумів їхню подібність, але лише BERT стабільно трактував такі формулювання як одне поняття. Це стало можливим завдяки контекстуальному аналізу, де модель здатна врахувати не лише лексичне значення слова, але й його оточення у реченні.

Ще одним важливим результатом було те, що BERT показав стійкість до орфографічних помилок. У реальних повідомленнях користувачі часто можуть помилятися в написанні слів. Наприклад: «квартера на Сихові», «житло в цетрі», «будинок біля вакзалу». Такі запити TF-IDF сприймав як невідомі терміни і майже завжди повертав нерелевантні результати. Word2Vec працював краще, але теж втрачав точність. Натомість BERT завдяки внутрішньому механізму субсловних токенів справлявся з такими варіаціями значно краще, правильно інтерпретуючи намір користувача.

У межах дослідження також було проаналізовано роботу системи на довгих запитах, що містять складні фрази і багатошаровий зміст. Наприклад, запит «хочу простору квартиру з новим ремонтом, окремими кімнатами та бажано ближче до парку або зони відпочинку; ціна до 600 доларів, але якщо дуже гарний варіант, можу трохи більше» є типовим для спілкування з реальними клієнтами. Під час тестів TF-IDF “розмивав” зміст такого запиту, оскільки значна кількість слів зменшувала значення ключових термінів. Word2Vec частково зберігав семантичну структуру, але неправильно інтерпретував умовні конструкції. Натомість BERT чітко визначав усі компоненти запиту: вид ремонту, форму планування, близькість до зелених зон, діапазон ціни та можливість коригування бюджету. Завдяки цьому система повертала максимально точні результати.

Практичні випробування чат-бота показали, що BERT не лише справляється з інтерпретацією складних запитів, але й здатний працювати в інтерактивному режимі, підтримуючи діалог і реагуючи на уточнення користувача. Наприклад, якщо після отримання результатів користувач уточнював «а можна щось ближче до

школи», система переобчислювала релевантність із урахуванням нової умови та оновлювала добірку. Це наближало роботу чат-бота до реальної взаємодії з брокером.

Загалом, результати дослідження свідчать, що трансформерні моделі, зокрема BERT, є найбільш перспективними для автоматизації пошуку нерухомості. Вони здатні інтерпретувати складні, нечіткі та багаторівневі запити, враховувати контекст, семантичні зв'язки, а також стилістичні особливості мовного оформлення. Отримані експериментальні результати дають підстави стверджувати, що поєднання модулів BERT із модульною архітектурою чат-бота забезпечує високий рівень точності та адаптивності, що робить систему придатною для реального застосування у брокерській діяльності та онлайн-платформах з нерухомості.

Висновки

Проведене дослідження дозволило отримати комплексне розуміння можливостей і ефективності застосування сучасних методів опрацювання природної мови у чат-ботах, призначених для автоматизації пошуку нерухомості. Розроблений у межах роботи прототип системи продемонстрував, що поєднання інструментів семантичного аналізу тексту з модульною архітектурою діалогового агента забезпечує значно вищу точність та релевантність результатів порівняно з традиційними підходами на основі ключових слів. Проведений експеримент, який охопив аналіз 520 оголошень та 75 природномовних запитів, підтвердив, що класичні моделі, зокрема TF-IDF, виявляються недостатньо придатними для задачі пошуку нерухомості через відсутність здатності враховувати контекст, синонімію та семантичну близькість між словами. Модель Word2Vec показала кращі результати, проте також виявила обмеження у випадках складних або нечітких запитів.

Найвищу ефективність продемонструвала трансформерна модель BERT, яка завдяки контекстуальному поданню слів змогла забезпечити значне підвищення точності пошуку. Значення Precision@5, що досягло рівня 0,83, свідчить про здатність моделі формувати коректні та релевантні добірки, навіть коли запит містить непрямі або недостатньо структуровані умови. Показник MRR = 0,79 додатково підтверджує стабільність роботи системи та високу ймовірність появи релевантних об'єктів на початкових позиціях пошукової видачі. Водночас середній час опрацювання запиту, який становив близько 112 мілісекунд, є цілком прийнятним для інтерактивної взаємодії у чат-боті, що дозволяє впроваджувати подібні рішення у реальних бізнес-процесах без суттєвих затримок.

Загалом результати дослідження засвідчують, що інтеграція сучасних NLP-моделей у процес пошуку нерухомості може відчутно підвищити рівень послуг і продуктивність роботи брокерів. Використання чат-бота як інструмента первинної взаємодії з клієнтом дозволяє значною мірою автоматизувати аналіз запитів, скоротити час на підбір варіантів, зменшити кількість нерелевантних результатів і забезпечити персоналізований підхід до кожного користувача. З технічної точки зору запропонована система є гнучкою, масштабованою та придатною для подальшої інтеграції з зовнішніми сервісами, такими як CRM, бази об'єктів або аналітичні платформи.

Новизна отриманих результатів зумовлена поєднанням порівняльного експерименту трьох NLP-підходів із практичним створенням і тестуванням діалогового агента, орієнтованого на специфіку ринку нерухомості. На відміну від існуючих оглядових досліджень, результати цієї роботи мають прикладний характер і ґрунтуються на реальних даних, що дозволяє робити висновки, безпосередньо придатні для використання в галузі.

Перспективи подальших досліджень пов'язуються з удосконаленням алгоритмів інтерпретації складних та багатокомпонентних запитів, розширенням функціональних можливостей чат-бота, впровадженням рекомендаційних механізмів на основі машинного навчання та прогнозних моделей, а також підтримкою багатомовних діалогових інтерфейсів. Окремим напрямом може стати розроблення мультиагентної архітектури, у якій окремі модулі відповідатимуть за пошук, оцінювання ринкових тенденцій, персоналізацію взаємодії з клієнтом і формування прогнозів вартості. Таким чином, результати проведеної роботи становлять основу для формування інтелектуальних систем підтримки нового типу брокерської діяльності у сфері нерухомості.

Література

1. Nosko V. The use of chatbots in business processes: Current trends and prospects // Business Inform. – 2020. – № 6. – P. 45–51.

2. Petrov A., Ivanenko O. Using NLP algorithms in chatbots for processing customer requests // System Research and Information Technologies. – 2019. – № 4. – P. 22–29.

3. Radziwill N. M., Benton M. C. Evaluating quality of chatbots and intelligent conversational agents // Journal of MultiDisciplinary Evaluation. – 2017. – Vol. 13, № 29. – P. 17–22.

4. Zhou L., Gao J., Li D., Shum H.-Y. The design and implementation of intelligent chatbots // ACM Computing Surveys. – 2020. – Vol. 53, № 5. – P. 1–38. – DOI: 10.1145/3405755.

5. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space // arXiv preprint. – 2013. – arXiv:1301.3781. – DOI: 10.48550/arXiv.1301.3781.

6. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies. – 2019. – P. 4171–4186. – DOI: 10.18653/v1/N19-1423.

7. Kravchenko I. Artificial intelligence in real estate: Challenges and opportunities // Economy and State. – 2021. – № 3. – P. 112–116.

8. NLP and transfer learning [Electronic resource] / Fast Forward Labs // Fast Forward Labs Technical Blog. – 2019. – Access mode: https://blog.fastforwardlabs.com/2019/08/28/nlp-and-transfer-learning.html.

References

1. Nosko V. The use of chatbots in business processes: Current trends and prospects // Business Inform. – 2020. – № 6. – P. 45–51.

2. Petrov A., Ivanenko O. Using NLP algorithms in chatbots for processing customer requests // System Research and Information Technologies. – 2019. – № 4. – P. 22–29.

3. Radziwill N. M., Benton M. C. Evaluating quality of chatbots and intelligent conversational agents // Journal of MultiDisciplinary Evaluation. – 2017. – Vol. 13, № 29. – P. 17–22.

4. Zhou L., Gao J., Li D., Shum H.-Y. The design and implementation of intelligent chatbots // ACM Computing Surveys. – 2020. – Vol. 53, № 5. – P. 1–38. – DOI: 10.1145/3405755.

5. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space // arXiv preprint. – 2013. – arXiv:1301.3781. – DOI: 10.48550/arXiv.1301.3781.

6. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2019. – P. 4171–4186. – DOI: 10.18653/v1/N19-1423.

7. Kravchenko I. Artificial intelligence in real estate: Challenges and opportunities // Economy and State. – 2021. – № 3. – P. 112–116.

8. NLP and transfer learning [Electronic resource] / Fast Forward Labs // Fast Forward Labs Technical Blog. – 2019. – Access mode: https://blog.fastforwardlabs.com/2019/08/28/nlp-and-transfer-learning.html.