УДК 004.912:811.161.2’282

ІЄРАРХІЧНА КЛАСИФІКАЦІЯ УКРАЇНСЬКИХ ДІАЛЕКТИЗМІВ ЗА ГЕОГРАФІЧНИМ ПОХОДЖЕННЯМ НА ОСНОВІ НЕЙРОННИХ МЕРЕЖ

Г. І. Дулішкович[1], О. В. Міца[2]

[1] Державний вищий навчальний заклад «Ужгородський національний університет», пл. Народна, 3, м. Ужгород, 88000, Україна, https://orcid.org/0000-0003-2493-2542, e-mail: dulgeoion@gmail.com
[2] Державний вищий навчальний заклад «Ужгородський національний університет», пл. Народна, 3, м. Ужгород, 88000, Україна, https://orcid.org/0000-0002-6958-0870, e-mail: alex.mitsa@uzhnu.edu.ua

Стаття присвячена розробці та оцінюванню ефективності системи автоматичної географічної локалізації українських діалектизмів із використанням нейромережевих методів. Запропонований підхід забезпечує визначення походження діалектних одиниць на трьох рівнях адміністративно-територіального поділу України: області, району та населеного пункту. Для аналізу та моделювання було використано корпус із 46 905 унікальних діалектизмів, сформований на базі краудсорсингової платформи «Інтерактивна мапа діалектів України» (dialectmap.org). Процес обробки та моделювання даних реалізовано у середовищі Python із використанням бібліотек PyTorch та FastAPI. Запропоновано гібридну архітектуру CharBiLSTM, що складається з двох паралельних символьних енкодерів для спільного опрацювання орфографічного запису слова та його фонетичної транскрипції. На першому кроці ієрархії модель здійснює класифікацію на рівні 15 областей України, досягаючи точності Top-3 97,36 % та Top-5 99,76 % на контрольній вибірці з 5 000 слів. Наступні рівні деталізації (район та населений пункт) реалізовано через каскадну фільтрацію пошукового індексу та пошук k найближчих сусідів (k-NN) за косинусною подібністю у 512-вимірному латентному просторі ознак. Точність на рівні району становить 51,18 %, а на рівні населеного пункту – 62,51 %, що зумовлено нерівномірністю емпіричного розподілу даних та наявністю опорних точок збору лінгвістичного матеріалу. Запропонований ієрархічний підхід з механізмом лінгвістичної пояснюваності є перспективним інструментом для автоматизації роботи діалектологів та інтеграції в сучасні геоінформаційні системи.

Ключові слова: *класифікація діалектизмів, ієрархічна класифікація, k-NN, інтерактивна карта діалектів, обробка природної мови, лінгвістична пояснюваність.*

Постановка проблеми. Українська мова володіє унікальним діалектним розмаїттям, що відображає складні історичні, культурні та етнографічні процеси формування регіональних спільнот. Проте в сучасних умовах глобалізації, урбанізації

та інтенсивної мовної уніфікації спостерігається поступове згасання діалектного мовлення, що робить задачу фіксації та збереження цього культурного надбання надзвичайно актуальною. Традиційні методи діалектології, засновані на ручному зборі даних під час польових експедицій та подальшому паперовому картографуванні, хоча й залишаються фундаментальними, мають істотні обмеження щодо швидкості обробки великих масивів інформації, масштабованості та можливостей глибокого статистичного аналізу.

Цифрова трансформація гуманітарних наук уможливила появу спеціалізованих веб-платформ, серед яких особливе місце посідає «Інтерактивна мапа діалектів України» (dialectmap.org), яка функціонує з 2020 року [1]. Платформа накопичує величезний масив георефернованих лінгвістичних даних, де кожен запис містить не лише орфографічний запис слова, а й його фонетичну транскрипцію, літературний відповідник та точні географічні координати населеного пункту побутування. На сьогодні вона містить понад 28 700 літературних слів та 38 900 діалектизмів, зібраних науковцями з різних регіонів України. За роки функціонування системи було зареєстровано 291 користувача, з яких 161 є активним, а середня кількість слів, внесених одним дослідником, становить 247 одиниць. Накопичення таких деталізованих даних відкриває можливості для застосування сучасних методів обробки природної мови та машинного навчання з метою автоматичної класифікації та локалізації діалектних одиниць.

Задача автоматичної географічної локалізації є нетривіальною через три ключові чинники. По-перше, значна кількість діалектизмів не має строго локалізованого характеру і може одночасно функціонувати в межах кількох суміжних областей або навіть різних діалектних груп. По-друге, фонетичні особливості слів хоча й містять цінні регіональні ознаки, не завжди корелюють із адміністративними кордонами. По-третє, наявні емпіричні дані характеризуються вираженим географічним дисбалансом: західні регіони України (зокрема, Закарпатська та Львівська області) представлені тисячами детальних записів, тоді як центральні, південні та східні області містять лише сотні записів. Подібні виклики, пов'язані з нерівномірністю класів та розмитістю меж, є загальною проблемою для задач комп'ютерної лінгвістики у слов'янських мовах. Це зумовлює потребу у створенні каскадних систем, здатних поєднувати нейромережеве моделювання з алгоритмами метричного пошуку для покрокового звуження пошукового простору.

Аналіз останніх досліджень та публікацій. Останніми роками методи машинного навчання все ширше інтегруються в завдання діалектології та лінгвістичного аналізу. Передумовою для цього стала розробка цифрових інструментів фіксації діалектів. У роботах дослідників детально висвітлено етнокультурний, освітній та науковий потенціал інтерактивних карт, які дозволяють залучати широке коло користувачів до збору лінгвістичного матеріалу за принципом краудсорсингу. Науковий аналіз текстових масивів української мови раніше успішно здійснювався за допомогою методу опорних векторів (SVM), зокрема для ідентифікації належності текстів до конкретних мас-медіа на основі їхніх лексичних особливостей [2, 3]. Ці дослідження підтвердили високу інформативність лексичного складу для розрізнення

джерел походження україномовного контенту та сформували методологічне підґрунтя для переходу до нейромережевого моделювання діалектного матеріалу.

Тісна взаємодія лінгвістів та фахівців з інформаційних технологій сприяла появі мобільних застосунків, заснованих на краудсорсинговому підході до збору мовних даних та побудови діалектних карт. Одним із таких рішень є мобільний додаток Dialäkt Äpp, який дає змогу користувачам фіксувати регіональні особливості швейцарських варіантів німецької мови та визначати ймовірне місце проживання користувача за результатами опитування, що складається з 16 запитань [4].

Подібний підхід реалізовано й у застосунку English Dialects, створеному Адріаном Лееманом спільно з дослідниками університетів Цюриха та Берна. Система аналізує вимову 26 слів і на основі отриманих відповідей встановлює найбільш імовірний різновид англійського діалекту. Результати відображаються у вигляді теплової карти, яка демонструє географічне поширення відповідних мовних особливостей [5].

Для моделювання послідовностей символів у лінгвістичних задачах класичним та перевіреним рішенням є рекурентні нейронні мережі, зокрема архітектура довгої короткострокової пам'яті (LSTM), запропонована С. Хохрайтером та Ю. Шмідхубером [6]. Завдяки здатності ефективно моделювати ліво-праві залежності та зберігати контекст на різних рівнях абстракції, LSTM демонструє стабільність у навчанні на вибірках середнього обсягу (до 50 000 прикладів), де сучасні трансформерні архітектури часто схильні до перенавчання. У задачах географічної локалізації текстів та класифікації діалектів використання символьних представлень є більш виправданим, ніж слівних, оскільки діалектні відмінності часто виражаються на рівні окремих фонем або суфіксів, які втрачаються при стандартній токенізації [7]. Незважаючи на значні успіхи в класифікації текстів, у сучасній комп'ютерній лінгвістиці практично відсутні інтегровані рішення для автоматичного ієрархічного аналізу окремих діалектних слів на кількох адміністративних рівнях (область, район, населений пункт). Традиційні класифікатори зазвичай обмежуються передбаченням лише одного рівня макрорегіону, що не задовольняє вимог детального лінгвістичного аналізу. Це вказує на необхідність розробки гібридних систем, що поєднують глибокі векторні представлення слів із каскадними процедурами пошуку у просторі ознак.

Мета статті. Метою статті є розробка, математичне обґрунтування та експериментальна оцінка інтегрованої автоматизованої системи ієрархічної географічної класифікації українських діалектизмів на основі поєднання двонапрямленої нейронної мережі символьного рівня (CharBiLSTM) та метричного алгоритму пошуку найближчих сусідів (*k*-NN).

Виклад основного матеріалу дослідження. Для навчання та тестування системи класифікації було залучено первинний масив даних із бази даних вебплатформи dialectmap.org. Оскільки вихідний набір даних формувався за допомогою краудсорсингу та експедиційних записів різних років, він містив певну кількість шумів, неповних метаданих та дублікатів. Процедура попереднього очищення даних включала видалення повторюваних записів за унікальною парою «діалектизм — населений пункт» для запобігання штучному зміщенню моделі,

ініціалізацію логічних фільтрів для виключення записів, що не мали точної географічної прив'язки або географічних координат, а також верифікацію та виправлення некоректних чи неповних фонетичних транскрипцій.

Після завершення очищення було сформовано репрезентативний корпус, який налічує 46 905 унікальних пар «діалектизм – транскрипція» з повними географічними метаданими. Дані охоплюють 15 областей України. Розподіл за областями є суттєво нерівномірним: Львівська область представлена найбільшою кількістю унікальних діалектизмів (понад 17 000), за нею йдуть Закарпатська (понад 12 000), Волинська та Івано-Франківська. Центральні та східні області представлені одиничними записами, тоді як центральні та східні області мають значно менше записів. Автори свідомо відмовилися від методів штучного балансування вибірки (таких як надлишкова вибірка або генерація синтетичних даних), щоб зберегти природний розподіл діалектного матеріалу та відобразити реальну інтенсивність побутування діалектизмів у різних зонах.

Кожен запис у корпусі має таку структуру: діалектизм (текстове представлення), фонетична транскрипція, літературний відповідник, область, район, населений пункт та географічні координати. Для безпосереднього навчання нейромережевого класифікатора використовуються лише текстові поля діалектизму та його транскрипції, тоді як адміністративні мітки нижчих рівнів залучаються на етапі ієрархічного пошуку. Розподіл даних на навчальну та тестову вибірки проведено у пропорції 80/20 із обов'язковою стратифікацією за областями, що гарантує збереження ідентичних пропорцій представлення регіонів в обох підмножинах. Обсяг контрольної тестової вибірки склав 5 000 слів.

В основі системи лежить двонапрямлена рекурентна мережа з довгою короткостроковою пам'яттю (BiLSTM) [6]. Вибір цієї архітектури зумовлений двома факторами. При розмірі корпусу менше 50 000 прикладів рекурентні мережі тренуються стабільніше порівняно з трансформерними архітектурами і менш схильні до перенавчання. Крім того, символьний рівень обробки (character-level) добре узгоджується з архітектурою LSTM: послідовність символів має виражену ліву-праву залежність, яку рекурентні мережі моделюють природно.

Модель складається з двох паралельних CharBiLSTM-енкодерів. Перший приймає посимвольне представлення діалектизму, другий обробляє його фонетичну транскрипцію. Наявність двох окремих енкодерів зумовлена тим, що орфографічне написання слова та його фонетична реалізація несуть різну, але взаємодоповнювальну інформацію про регіональне походження. Наприклад, транскрипція може відображати характерні для певного регіону фонетичні явища (як-от закарпатське укання або поліське дифтонгічне вимовляння), які не завжди відображені в орфографічному записі.

Кожен енкодер містить шар символьних вкладень (embedding) розмірністю 64 та двошаровий BiLSTM із прихованою розмірністю 128. На виході кожний енкодер формує 256-вимірний вектор (конкатенація останніх прихованих станів прямого і зворотного напрямків). Об'єднання виходів двох енкодерів дає 512-вимірне представлення діалектизму (рис. 1).

Рис. 1. Архітектура моделі ієрархічної класифікації діалектизмів

Блок класифікації складається з повнозв›язного шару (512 → 128), функції активації ReLU, шару Dropout (0.3) для регуляризації та вихідного шару (128 → 15 областей). Загальна кількість параметрів моделі становить 1,27 млн.

Навчання проводилось протягом 30 епох з оптимізатором Adam [8] (learning rate 0.001) та функцією втрат CrossEntropyLoss. Розмір міні-батчу становив 128. Для запобігання перенавчанню, окрім Dropout, використовувалось раннє припинення навчання (early stopping): процес зупинявся, якщо метрика Top-3 точності на валідаційній вибірці не покращувалась протягом 5 послідовних епох. Зберігалась модель з найкращим значенням цієї метрики.

Ключовою особливістю розробленої системи є каскадна процедура класифікації, яка працює у три послідовні кроки, поступово звужуючи географію від обласного до локального рівня.

На першому кроці нейронна мережа обчислює ймовірність належності діалектизму до кожної з 15 областей на основі його 512-вимірного вектора. Для цього використовується навчений класифікатор із softmax-активацією на виході. Результатом є ранжований список областей з відповідними ймовірностями. Для кожного прогнозу обчислюється показник впевненості (максимальна ймовірність); якщо він нижчий за заданий поріг, система сигналізує дослідникові про необхідність додаткової верифікації.

На другому кроці використовується пошук k найближчих сусідів (*k*-NN) [9] серед усіх 46 905 проіндексованих слів. Попередньо для кожного слова в індексі обчислено 512-вимірний вектор за допомогою тієї ж моделі. При запиті індекс

фільтрується за визначеною на першому кроці областю, і серед відфільтрованих слів знаходяться k=50 найближчих сусідів за косинусною подібністю. Бали подібності агрегуються за районами і нормалізуються, формуючи ранжований список районів.

Третій крок аналогічний другому: індекс додатково фільтрується за визначеним районом, і k-NN-пошук серед решти слів визначає найбільш імовірні населені пункти.

Важливою характеристикою системи є механізм пояснюваності. На кожному кроці ієрархії дослідник бачить перелік найбільш подібних слів з індексу разом із їх географічним розташуванням та показником косинусної подібності. Цей перелік дозволяє діалектологу верифікувати прогноз на основі лінгвістичних аналогій: якщо серед найближчих сусідів переважають слова з певного регіону, це підтверджує обґрунтованість прогнозу. Ми вважаємо цей механізм не менш важливим, ніж сам прогноз, оскільки в діалектологічних дослідженнях довіра до результату суттєво залежить від можливості його інтерпретації.

Якість моделі оцінено на тестовій вибірці з 5 000 діалектизмів. Для порівняння було реалізовано три версії моделі, які відрізняються архітектурою та підходом до класифікації (див. табл. 1).

Таблиця 1

Порівняння точності моделей на обласному рівні

Модель	Top-1	Top-3	Top-5
V8 (базова, FC)	65,71 %	90,57 %	98,31 %
V9 (BiLSTM)	75,62 %	93,69 %	99,00 %
V10 (BiLSTM + k-NN)	74,96 %	97,36 %	99,76 %

Модель V8 є базовою архітектурою, що ґрунтується на використанні повнозв'язних нейронних шарів і не передбачає рекурентної обробки символьних послідовностей. Модель V9 використовує архітектуру Char-BiLSTM, аналогічну моделі V10, однак виконує класифікацію лише на рівні областей без застосування ієрархічного механізму пошуку. Модель V10 поєднує Char-BiLSTM-класифікатор із каскадним k-NN-пошуком, що забезпечує послідовну локалізацію діалектизмів на різних рівнях адміністративно-територіального поділу.

Як видно з рисунку 2, модель V9 має дещо вищий Top-1 (75,62% проти 74,96% у V10). Незначне зниження Top-1 у V10 пояснюється тим, що k-NN-агрегація може перерозподіляти ваги між областями. Проте V10 суттєво перевершує V9 у Top-3 (97,36 % проти 93,69 %) та Top-5 (99,76 % проти 99,00 %). Приріст порівняно з базовою моделлю V8 становить +9,25 в.п. для Top-1, +6,79 в.п. для Top-3 та +1,45 в.п. для Top-5. Для дослідницьких задач, де важливо знайти правильну відповідь серед кількох кандидатів, ієрархічний підхід з k-NN ефективніший за «чисту» класифікацію.

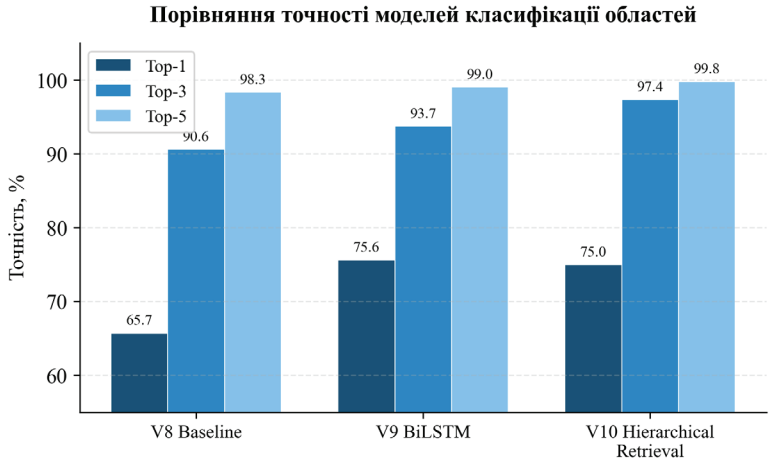

Рис. 2. Порівняння точності моделей V8, V9, V10

На нижчих рівнях ієрархії показники очікувано нижчі (рис. 3). На рівні району точність складає 51,18 %, на рівні населеного пункту – 62,51 %. З огляду на велику кількість класів (близько 150 районів і понад 3 000 населених пунктів) та нерівномірний розподіл даних, ці результати є прийнятними. Зростання точності на рівні населених пунктів порівняно з районами пояснюється тим, що в межах одного району зазвичай домінує один населений пункт з найбільшою кількістю записів.

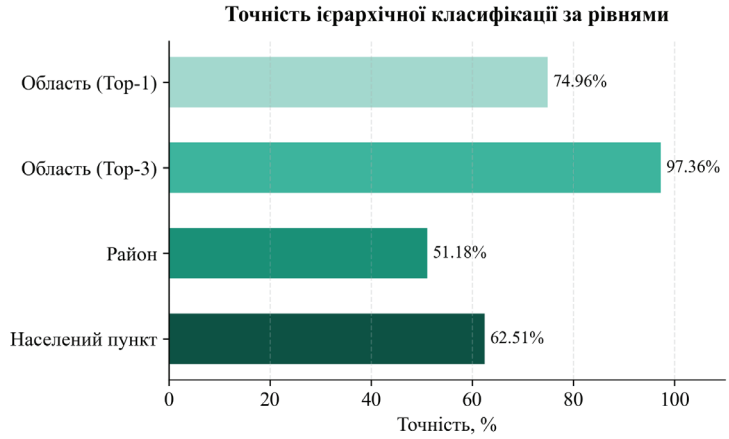

Рис. 3. Точність ієрархічної класифікації за рівнями

Аналіз помилок показав характерні закономірності. Найчастіше модель плутає суміжні області, що належать до одного діалектного масиву: наприклад, Закарпатську та Львівську (обидві належать до південно-західного наріччя). Рідкісні діалектизми з областей, слабко представлених у корпусі, найчастіше помилково приписуються до домінуючих областей. Ці спостереження вказують на перспективність розширення корпусу саме за рахунок центральних та східних областей.

Для практичного використання моделі розроблено REST API мовою Python на базі фреймворку FastAPI [10]. Вибір FastAPI зумовлений його асинхронною архітектурою на основі ASGI, автоматичною генерацією інтерактивної документації (Swagger UI та ReDoc) та інтеграцією з Pydantic для валідації вхідних і вихідних даних. Структура API відображає ієрархічну природу класифікації та пропонує кілька режимів роботи.

Інференс здійснюється в PyTorch [11] без обчислення градієнтів (torch.no_grad), що дозволяє мінімізувати споживання пам'яті та час відповіді. Модель та пошуковий індекс завантажуються в пам›ять при старті сервера через механізм lifespan у FastAPI. Валідація вхідних даних (обмеження на довжину слова, перевірка існування вказаної області чи району) реалізована через Pydantic-моделі з типізованими полями та обмеженнями.

API інтегровано з існуючим бекендом платформи dialectmap.org на Ruby on Rails через внутрішні HTTP-запити. Для зручності користувачів створено веб-інтерфейс класифікатора у вигляді односторінкового додатку, який дозволяє ввести діалектизм та отримати візуалізацію ієрархічного прогнозу на трьох рівнях із переліком подібних слів [12].

Висновки. У межах проведеного дослідження успішно вирішено задачу автоматизованої ієрархічної класифікації українських діалектизмів за географічним походженням. Запропоноване поєднання двонапрямленої рекурентної мережі символьного рівня CharBiLSTM з каскадним *k*-NN-пошуком продемонструвало високу точність на макрорегіональному рівні, забезпечивши показники Top-3 точності 97,36% та Top-5 точності 99,76% на тестовій вибірці з 5 000 слів.

Порівняльний аналіз трьох конфігурацій моделей підтвердив наукову доцільність каскадної архітектури: хоча модель V10 незначно поступається «чистій» класифікації за метрикою Top-1 через ефект метричного згладжування, вона суттєво перевершує її у виявленні множини потенційних регіонів побутування слова. Інтегрований механізм лінгвістичної пояснюваності через демонстрацію подібних слів та їхніх географічних координат забезпечує прозорість роботи системи для фахівців-діалектологів.

Практична цінність дослідження полягає в реалізації REST API на базі FastAPI, що дозволило безшовно інтегрувати інтелектуальний класифікатор у структуру діючої платформи dialectmap.org, зробивши його доступним як для академічних досліджень, так і для широкого загалу.

Серед ключових обмежень розробленої системи варто виділити виражену нерівномірність географічного представлення навчальних даних у вихідному корпусі, відсутність уніфікованої транскрипції для значної частини зафіксованого матеріалу, а також фокусування аналізу виключно на лексичному рівні без урахування ширшого синтаксичного контексту чи морфологічних парадигм.

Основними напрямками подальших досліджень вбачаються розширення навчального корпусу за рахунок наповнення бази даних новими записами з центральних та східних регіонів України для ліквідації географічного дисбалансу, апробація сучасних архітектур трансформерного типу для покращення моделювання довгих

символьних залежностей, перехід до регресійного прогнозування географічних координат замість дискретної класифікації, а також глибша інтеграція інструментів штучного інтелекту безпосередньо в інтерфейс інтерактивної мапи діалектів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Міца О.В., Шумицька Г.В., Шаркань В.В., Венжинович Н.Ф., Дулішкович Г.І. (2022). Інтерактивна карта діалектів як інструмент фахової підготовки студентів філологічних спеціальностей. Інформаційні технології і засоби навчання, 88(2), 126–138. https://doi.org/10.33407/itlt.v88i2.4787.
2. Lupei, M., Mitsa, O., Sharkan, V., Vargha, S., & Gorbachuk, V. (2022). The identification of mass media by text based on the analysis of vocabulary peculiarities using support vector machines. In Proceedings of the 2022 International Conference on Smart Information Systems and Technologies (SIST) (pp. 1–6). IEEE. https://doi.org/10.1109/SIST54437.2022.9945774.
3. Lupei, M., Mitsa, O., Sharkan, V., Vargha, S., & Lupei, N. (2023). Analyzing Ukrainian media texts by means of support vector machines: Aspects of language and copyright. In Advances in Computer Science for Engineering and Education VI (Lecture Notes on Data Engineering and Communications Technologies, Vol. 181, pp. 173–182). Springer. https://doi.org/10.1007/978-3-031-36118-0_16.
4. Leemann, A., Kolly, M.-J., Purves, R., Britain, D., & Glaser, E. (2016). Crowdsourcing language change with smartphone applications. PLoS ONE, 11(1), Article e0143060. https://doi.org/10.1371/journal.pone.0143060.
5. Leemann, A., Kolly, M.-J., & Britain, D. (2018). The English Dialects App: The creation of a crowdsourced dialect corpus. Ampersand, 5, 1–17.
6. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
7. Macko, D., Moro, R., Uchendu, A., Lucas, J., Yamashita, M., Pikuliak, M., & Šimko, J. (2023). MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 9960–9987). https://doi.org/10.18653/v1/2023.emnlp-main.616.
8. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6980.
9. Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964.
10. Ramírez, S. (2018). FastAPI: Modern, fast web framework for building APIs with Python. https://fastapi.tiangolo.com.
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, 32, 8024–8035.
12. Lupei, M., Shlahta, M., Mitsa, O., Horoshko, Y., Tsybko, H., & Gorbachuk, V. (2022). Development of an interactive map within the implementation of actual state and public directions. In Proceedings of the 2022 12th International Conference on Advanced Computer

Information Technologies (ACIT) (pp. 384–387). IEEE. https://doi.org/10.1109/ACIT54803.2022.9913191.

REFERENCES

1. Mitsa, O. V., Shumytska, H. V., Sharkan, V. V., Venzhynovych, N. F., & Dulishkovych, H. I. (2022). Interaktyvna karta dialektiv yak instrument fakhovoi pidhotovky studentiv filolohichnykh spetsialnostei [Interactive dialect map as a tool for professional training of philology students]. *Informatsiini Tekhnolohii i Zasoby Navchannia, 88*(2), 126–138. https://doi.org/10.33407/itlt.v88i2.4787.
2. Lupei, M., Mitsa, O., Sharkan, V., Vargha, S., & Gorbachuk, V. (2022). The identification of mass media by text based on the analysis of vocabulary peculiarities using support vector machines. In *Proceedings of the 2022 International Conference on Smart Information Systems and Technologies (SIST)* (pp. 1–6). IEEE. https://doi.org/10.1109/SIST54437.2022.9945774.
3. Lupei, M., Mitsa, O., Sharkan, V., Vargha, S., & Lupei, N. (2023). Analyzing Ukrainian media texts by means of support vector machines: Aspects of language and copyright. In *Advances in Computer Science for Engineering and Education VI* (Lecture Notes on Data Engineering and Communications Technologies, Vol. 181, pp. 173–182). Springer. https://doi.org/10.1007/978-3-031-36118-0_16.
4. Leemann, A., Kolly, M.-J., Purves, R., Britain, D., & Glaser, E. (2016). Crowdsourcing language change with smartphone applications. *PLoS ONE, 11*(1), Article e0143060. https://doi.org/10.1371/journal.pone.0143060.
5. Leemann, A., Kolly, M.-J., & Britain, D. (2018). The English Dialects App: The creation of a crowdsourced dialect corpus. *Ampersand, 5*, 1–17.
6. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
7. Macko, D., Moro, R., Uchendu, A., Lucas, J., Yamashita, M., Pikuliak, M., & Šimko, J. (2023). MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9960–9987). https://doi.org/10.18653/v1/2023.emnlp-main.616.
8. Kingma, D. P., & Ba, J. (2015). *Adam: A method for stochastic optimization*. Proceedings of the 3rd International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6980.
9. Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964.
10. Ramírez, S. (2018). *FastAPI: Modern, fast web framework for building APIs with Python*. https://fastapi.tiangolo.com.
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems, 32*, 8024–8035.
12. Lupei, M., Shlahta, M., Mitsa, O., Horoshko, Y., Tsybko, H., & Gorbachuk, V. (2022). Development of an interactive map within the implementation of actual state and public directions. In *Proceedings of the 2022 12th International Conference on Advanced Computer Information Technologies (ACIT)* (pp. 384–387). IEEE. https://doi.org/10.1109/ACIT54803.2022.9913191.

doi: 10.32403/1998-6912-2026-1-72-170-180

HIERARCHICAL CLASSIFICATION OF UKRAINIAN DIALECT WORDS BY GEOGRAPHIC ORIGIN USING NEURAL NETWORKS

H. I. Dulishkovych[1], O. V. Mitsa[2]

[1] Uzhhorod National University, 3 Narodna Sq., Uzhhorod, 88000, Ukraine, https://orcid.org/0000-0003-2493-2542, e-mail: dulgeoion@gmail.com
[2] Uzhhorod National University, 3 Narodna Sq., Uzhhorod, 88000, Ukraine https://orcid.org/0000-0002-6958-0870, e-mail: alex.mitsa@uzhnu.edu.ua

The article is devoted to the development and evaluation of an automatic geographical localization system for Ukrainian dialect words using neural network methods. The proposed approach enables the identification of the origin of dialectal units at three levels of Ukraine's administrative-territorial division: region (oblast), district (raion), and settlement. For data analysis and modeling, a corpus of 46,905 unique dialect words was utilized, compiled from the crowdsourcing platform «Interactive Map of Ukrainian Dialects» (dialectmap.org). The data processing and modeling pipeline was implemented in Python using the PyTorch and FastAPI frameworks. A hybrid Char-BiLSTM architecture is proposed, consisting of two parallel character-level encoders that jointly process the orthographic representation of a word and its phonetic transcription. At the first stage of the hierarchy, the model performs classification among 15 Ukrainian regions, achieving Top-3 accuracy of 97.36% and Top-5 accuracy of 99.76% on a test set of 5,000 words.The subsequent levels of localization (district and settlement) are implemented through cascading index filtering and k-nearest neighbors (k-NN) search based on cosine similarity in a 512-dimensional latent feature space. The accuracy at the district level reaches 51.18%, while the settlement-level accuracy is 62.51%. These results are influenced by the uneven empirical distribution of the data and the presence of linguistic data collection anchor points. The proposed hierarchical approach, combined with a mechanism of linguistic explainability, represents a promising tool for automating dialectological research and integrating dialect localization capabilities into modern geographic information systems.

Keywords: *dialect classification, hierarchical classification, k-NN, Interactive Map of Ukrainian Dialects, natural language processing, linguistic explainability.*

Стаття надійшла до редакції 18.05.2026.
Submitted: 18.05.2026.
Прийнято до друку: 22.05.2026.
Accepted: 22.05.2026.
Опубліковано: 30.05.2026.
Published: 30.05.2026.