doi: 10.32620/oikit.2025.104.11

УДК 004.8:371.3:81’33

М. В. Шаповал, А. Г. Чухрай, Л. В. Мандрікова

Інтерактивна модель навчання критичному мисленню через ШІ-діалог

Національний аерокосмічний університет «Харківський авіаційний інститут»

У статті представлено міждисциплінарне дослідження, присвячене розробці інтерактивного діалогового тренажера, що використовує штучний інтелект для формування навичок критичного мислення у користувачів. Центральна ідея роботи полягає у зміні парадигми: не людина навчає ШІ, а ШІ, адаптуючись до користувача, виступає як тренер та співрозмовник, моделюючи різні типи маніпулятивної мовленнєвої поведінки. Тренажер має форму щоденного діалогу, що імітує реальні комунікативні ситуації, з метою поступового формування здатності користувача розпізнавати дезінформацію, приховані директиви, логічні помилки та психологічні впливи.

Концепція тренажера натхненна ідеями С. І. Хаякави, викладеними в його фундаментальній праці "Мова в дії" [6], де проаналізовано типові мовні маніпуляції, що знижують якість мислення. Саме цей концептуальний апарат став основою сценарного модулю тренажера, де кожен тип впливу — як апеляції до емоцій, авторитету, чи хибних дихотомій — відтворюється за допомогою відповідних алгоритмів генерації мови.

У технологічній частині дослідження проаналізовано можливості використання сучасних мовних моделей (BERT, XLNet, Bi-LSTM) у контексті освітньої діалогової взаємодії. Зокрема, розглянуто, як моделі класифікації дезінформації можуть бути адаптовані до навчального процесу з урахуванням психолінгвістичного профілю користувача. Запропоновано гібридну архітектуру системи з модулями адаптації, тематичного моделювання, зворотного зв’язку, семантичного аналізу й підтримки довготривалої взаємодії.

Стаття спрямована на створення основи для розробки повноцінного цифрового інструмента — тренажера, здатного не лише виявляти дезінформацію, а й формувати культуру усвідомленого мовлення та сприйняття інформації. Запропонована система може бути інтегрована в освітні платформи, зокрема для індивідуалізованого навчання студентів, журналістів, викладачів та інших професій, що потребують критичного аналізу інформації. Такий підхід сприяє не лише розвитку когнітивних навичок, а й формуванню стійкості до інформаційних впливів у цифрову епоху.

Ключові слова: критичне мислення; штучний інтелект; дезінформація; мовні маніпуляції; BERT; XLNet; тренажер; діалогова система; Хаякава; освітні технології.

Вступ

В епоху інформаційних війн і цифрової трансформації питання протидії дезінформації постає особливо гостро. Сучасне медіасередовище, за влучним висловом Маршалла Маклюена, перетворилося на «продовження нервової системи людини» [5]. У цьому контексті здатності критично мислити й усвідомлено вживати мову набувають першорядного значення. Як зазначав С. І. Хаякава, уміння *бачити крізь слова* – не привілей фахівців, а необхідна навичка кожного, хто живе в суспільстві, насиченому символами [6]. Його книга «*Мова в дії*» надихнула дане дослідження [6], поставивши собі за мету озброїти людину засобами протистояння вербальним маніпуляціям і «словесному гіпнозу» [6].

Переважна більшість сучасних робіт із боротьби з дезінформацією сфокусована на автоматичному виявленні фейків. Розробляються алгоритми класифікації текстів на правдиві та неправдиві, фільтрації маніпулятивних

повідомлень тощо. – тобто, по суті, автоматичні «щити» в інформаційному просторі. Для цих завдань сьогодні успішно застосовуються моделі глибокого навчання: трансформери (наприклад, BERT, XLNet), рекурентні нейромережі (Bi-LSTM), механізми уваги (attention), тематичне моделювання та ансамблеві методи. Так, у нещодавній дисертації запропоновано комплексний фреймворк, що комбінує BERT і Bi-LSTM для одночасного виявлення дезінформації та тематичного аналізу тексту [1]. Ці методи демонструють високу точність при автоматичному аналізі постів у соцмережах і новинних статей [1].

Однак протидія маніпуляціям може здійснюватися не лише через автоматичну класифікацію, а й через навчання користувачів. Замість того щоб просто позначати контент як фейковий, перспективним підходом є інтерактивне навчання людини розпізнавати хитрощі самостійно. Для цього пропонується інтелектуальний діалоговий тренажер: ШІ-система, яка веде з користувачем реалістичну розмову, наповнюючи її прихованими мовленнєвими та смисловими маніпуляціями. Мета - в умовах, наближених до реального спілкування, розвинути в користувача критичне мислення та навички виявлення хибних аргументів. У такій системі ШІ виступає в ролі вчителя, адаптуючись під стиль і аргументацію того, кого навчають, а не навпаки. Нижче розглянуто, як перелічені вище моделі та підходи можуть бути адаптовані під подібний навчальний діалог, які обмеження при цьому виникають і які архітектурні рішення можливі.

Ми пропонуємо саме такий підхід: інтелектуальний діалоговий тренажер, у якому ШІ-система веде з користувачем реалістичну розмову, поволі наповнюючи її прихованими мовленнєвими та смисловими маніпуляціями. Завдання - в умовах, наближених до реального спілкування, розвинути у користувача критичне мислення і навички виявлення помилкових аргументів. У цій системі ШІ виступає в ролі наставника, підлаштовуючись під стиль і логіку співрозмовника. Підхід співзвучний ідеям Хаякави про необхідність прищеплювати людям «семантичну грамотність» і здатність відрізняти карту (слова) від території (реальності) [6]. Нижче розглянуто, як перераховані моделі та методи можуть бути адаптовані під подібний навчальний діалог, з якими обмеженнями ми зіткнемося і які архітектурні рішення можливі.

1. Традиційні моделі аналізу дезінформації та їхні можливості

Перш ніж перейти до діалогової системи, важливо зрозуміти, як саме перераховані методи працюють у завданні автоматичного аналізу тексту і виявлення фейків. У дослідженнях з виявлення дезінформації ці підходи проявили себе таким чином:

Трансформери (BERT, XLNet) - потужні мовні моделі, що контекстуально кодують текст. BERT (Bidirectional Encoder Representations from Transformers) опрацьовує речення цілком в обох напрямках, чудово вловлюючи сенс і тонкі нюанси формулювань. XLNet - авторегресійна модель з перестановочним механізмом навчання, ефективно передбачає послідовності слів. Було показано, що трансформерні моделі (включно з іншими варіантами на кшталт ELECTRA і ALBERT) досягають дуже високих показників точності при виявленні фейкових повідомлень [2]. У дослідженні показано, що трансформерні моделі (включно з, наприклад, ELECTRA, LAMBERT) досягали дуже високих метрик при виявленні фейків [1]. BERT і XLNet застосовували для класифікації постів у соцмережах на пропагандистські/маніпулятивні чи ні, даючи високу точність (F1-бал до ~95% на окремих датасетах) [1] . Це означає, що трансформери здатні розпізнавати,

наприклад, **апеляції до емоцій** (через виявлення емоційно навантажених формулювань) або приховані повідомлення у фразах, які на перший погляд виглядають нейтрально. Їхня сила - в умінні виявляти приховані смислові зв'язки та контекстуальні натяки, які часто присутні в маніпулятивних повідомленнях.

Рекурентні нейромережі (Bi-LSTM) - двоспрямовані мережі довгої короткострокової пам'яті, що послідовно опрацьовують текст з урахуванням як попереднього, так і наступного контексту. Bi-LSTM добре вловлюють послідовність аргументації, ланцюжки причинно-наслідкових тверджень - що корисно при аналізі логіки викладу. Раніше їх широко застосовували для аналізу послідовностей, і навіть зараз Bi-LSTM можуть доповнювати трансформери. Наприклад, згаданий фреймворк об'єднує BERT і Bi-LSTM: трансформер витягує глибокі ембеддінги тексту (векторні уявлення слів у багатовимірному просторі), а Bi-LSTM інтегрує лінійні та нелінійні залежності в послідовності [1]. Такий гібрид дозволив підвищити точність класифікації, оскільки поєднує сильні сторони обох архітектур. RNN/LSTM вловлюють структуру міркувань і причинно-наслідкові ланцюжки - це допомагає виявляти **логічні невідповідності** в тексті. У контексті фейкових новин гібридні моделі (наприклад, CNN+LSTM, Bi-GRU тощо) демонстрували високу ефективність, особливо в комбінації з методами балансування даних і тонкого налаштування моделей [1][4].

Механізми уваги (attention) – дають змогу фокусуватися на найбільш значущих фрагментах тексту. Вони застосовуються і всередині трансформерів, і в гібридних схемах (наприклад, Attention-BiLSTM). У дисертації Падалко Г. А. [1] представлено представлено модифікацію Attention-Based Bi-LSTM, де увага виявляє ключові слова та речення, важливі для класифікації. Багатоголові механізми уваги (multi-head attention) застосовуються у тематичному моделюванні: вони допомагають чітко виокремити тематичні кластери в ембеддінгах - векторних представленнях слів. допомогло чіткіше розділити тематичні кластери в даних [1]. Інакше кажучи, механізми уваги (attention) дає як краще розуміння сенсу (за рахунок виділення важливих деталей), так і підвищує інтерпретованість - можна побачити, на що модель звертає увагу під час винесення рішення. Такі системи здатні підкреслювати **емоційні стимулюючі слова**, які викликають емоційний відгук. По суті, це алгоритмічне наближення до того, чому Хаякава вчив чинити самостійно - бачити **емоційні маркери** [6].

Тематичне моделювання – вилучення тем у корпусі текстів (наприклад, за допомогою LDA або BERTopic, нейромережевого методу тематичного аналізу, що поєднує векторні представлення з кластеризацією) [13]. У завданні виявлення дезінформації тематичний аналіз корисний для розуміння, про що йдеться в потенційно фейковому тексті: політика, здоров'я, економіка тощо. У зазначеному дослідженні тема-моделювання інтегрована з виявленням фейків: запропоновано підхід, що об'єднує класифікацію та тематичний розбір, щоб не тільки визначити факт маніпуляції, а й дати її контекст [1]. Застосування сучасних нейромережевих методів з урахуванням уваги дало змогу чіткіше ділити тексти на теми навіть у змішаних наборах даних [1]. Таким чином, модель може не просто сказати «неправда», а розуміти, в якій галузі знань ця неправда, що допомагає при подальшій інтерпретації. Тематичний аналіз дає змогу пояснити користувачеві: чому саме тут використовується та чи інша хитрість, який наративний сюжет прихований за словами. Наприклад, виявивши, що текст апелює до теми безпеки, модель вкаже, що узагальнення на кшталт «суспільство вироджується» насправді експлуатує страх і належить до наративу кризи.

Ансамблеві методи – комбінація декількох моделей/алгоритмів (наприклад, BERT + LSTM + дерева рішень), що підвищує надійність. В аналізі соціальних мереж ансамблі часто використовують для підвищення стійкості: різні моделі доповнюють одна одну, згладжуючи помилки. У дисертації Падалко Г.А. [1] зазначено, що ансамблі, які враховують дисбаланс класів і адаптуються під нові дані, поліпшили якість класифікації фейкових повідомлень. Наприклад, можна об'єднати висновки кількох класифікаторів (скажімо, BERT, LSTM і метод на вирішальних деревах) - рішення за більшістю або зважене усереднення найчастіше точніше за будь-який одиночний класифікатор. Такі ансамблі корисні в контексті тренажера: вони забезпечують різноманітність мовленнєвих стратегій і забезпечують зворотний зв'язок за кількома осями аналізу. Як писав Хаякава, навіть точний аналіз не замінює розуміння мови. «Навіть найнебезпечніші слова не звучать як злоба - вони звучать як істини» [6]. Тому в основі ефективного навчання - не проста фільтрація «поганих» слів, а розвиток уміння помічати, рефлексувати та порівнювати різні висловлювання.

2. Інтелектуальний тренажер у формі діалогу

Принципова відмінність пропонованої нами системи від наявних алгоритмів аналізу дезінформації полягає в тому, що вона не просто класифікує повідомлення як правдиве або неправдиве, а активно залучає користувача до семіотично насиченого діалогу. У цьому діалозі штучний інтелект (ШІ) виконує роль партнера, а іноді й провокатора. Мета - не дати остаточного вердикту, а навчити саму людину бачити, розпізнавати, осмислювати маніпуляції в мові.

Такий підхід близький до методики Сократичного діалогу - коли вчитель не дає прямих відповідей, а навідними запитаннями підштовхує учня до самостійних висновків. У нещодавньому проєкті TITAN був створений чат-бот, що стимулює критичне мислення щодо новин саме через запитання, а не через готові відповіді [7]. Там ШІ відстежував ознаки дезінформації в тексті новини та ставив користувачеві послідовність запитань (уточнення, виявлення припущень, пошук доказів і альтернатив тощо), щоб той сам дійшов висновку про надійність інформації [7]. Цей приклад показує принцип: не говорити напряму, де брехня, а вчити людину самій її бачити. У нашому випадку тренажер може поводитися більш різноманітно, не тільки запитувати, а й стверджувати, тобто сперечатися, щоб максимально наблизити спілкування до реальних умов.

Діалоговий формат висуває нові вимоги до моделей: потрібна генерація осмислених відповідей, підтримання контексту в кількох репліках, розуміння реакцій користувача. Крім того, система має вміти адаптуватися - якщо користувач, скажімо, навів контраргумент, ШІ має підлаштуватися: або посилити натиск, змінити тактику маніпуляції, або (на певному етапі навчання) визнати маніпулятивний прийом і пояснити його. Нижче ми розглянемо, наскільки придатні описані раніше моделі для розв'язання цих завдань діалогу, і як їх можна вбудувати в архітектуру тренажера.

Такий тренажер моделює умови реального спілкування: користувач отримує репліки, що містять емоційні заклики, софізми, пересмикування чи риторичні пастки, а потім має спробувати їх розпізнати й аргументувати своє сприйняття. Це інтерактивна вправа на уважність до мови, стилю, інтонацій, підтекстів. Подібний підхід ми називаємо діалоговим, навчальним, «сократичним». Він натхненний не тільки античною педагогікою, а й сучасними поглядами на мову **як форму дії** за визначенням Хаякави [6].

У своїй книзі Хаякава підкреслює, що слова «живуть у контексті» і що суперечка - це не завжди пошук істини, а часто - боротьба за інтерпретацію. Він наводить приклад фрази «він вчинив по-людськи», яка може мати як позитивне, так і негативне значення – залежно від контексту, інтонації, позиції мовця [6]. Тренажер використовує це спостереження: система навмисно генерує багатозначні, контекстуально заряджені висловлювання і спостерігає, як їх інтерпретує користувач. Це не тест - це тренування інтерпретації, вміння ставити собі запитання: «що тут насправді сказано?».

У пропонованій системі ШІ-співрозмовник може приймати різні *комунікативні ролі*: агресивного опонента, іронічного співрозмовника, нав'язливого переконувача, конспіролога тощо. Кожна роль тренує розпізнавання різних типів маніпуляцій. Сценарії діалогу навмисно включають різноманітні хитрощі: **апеляцію до емоцій** («Хіба ти не дбаєш про свою сім'ю?»), **апеляцію до авторитету** («Вчені довели, що...**»), **хибні дихотомії** (пропонування вибору лише з двох крайніх варіантів), **широкі хибні узагальнення**, **приховані директиви** чи заклики, завуальовані під нейтральні твердження, **конотативні пастки** з використанням емоційно заряджених слів, а також спотворення чи селективне використання фактів і статистики. Користувач при цьому не отримує прямої вказівки «це брехня» - він має сам навчитися її розпізнавати. ШІ лише ненав'язливо підштовхує до спостереження, аналізу, уточнюючих запитань.

Таким чином, ШІ перетворюється на дидактичне середовище. Він веде діалог, накопичує дані про реакцію користувача, аналізує типові помилки і, в ідеалі, підлаштовує наступні сесії під найвразливіші аспекти сприйняття. Це не тільки навчає виявляти брехню, а й розвиває метамовну рефлексію - здатність помічати, як саме мова впливає на свідомість. Саме таку здатність Хаякава вважав фундаментом «мовної просвіти»: «лише той, хто вміє відрізняти між рівнем факту і рівнем емоції, здатен вибудувати ясну думку» [6].

3. Застосування трансформерів (BERT, XLNet) у діалоговій системі

Для побудови реалістичного діалогу, в якому користувач стикається з різноманіттям мовленнєвих маніпуляцій, необхідні мовні моделі, здатні не тільки розуміти, а й відтворювати людське мовлення з високим ступенем достовірності. BERT і XLNet являють собою найперспективніший клас архітектур, здатний виконати обидві ці задачі в рамках пропонованого тренажера.

Аналітична роль BERT (Bidirectional Encoder Representations from Transformers) Модель BERT має можливість двобічного аналізу контексту і може застосовуватися в системі як інтерпретатор користувацьких реплік. Вона класифікує кожне повідомлення користувача за низкою ознак: чи є в ньому індикатори розпізнання маніпуляції, яке емоційне забарвлення відповіді, ступінь згоди з тезою ШІ тощо. Наприклад, BERT-модель могла б визначати, чи повівся користувач на помилковий аргумент або, навпаки, засумнівався в ньому. Це особливо важливо для розуміння реакції: так система дізнається, помітив користувач емоційний прийом чи прийняв приховану аксіому на віру. На основі такої класифікації BERT оцінює, наскільки успішно людина розпізнала маніпуляцію, і які елементи тексту викликали нерозуміння або емоційний відгук. Крім того, модуль на основі BERT може виступати частиною механізму зворотного зв'язку - пояснюючи після спілкування, **що саме** в тексті було потенційно маніпулятивним і чому (наприклад, вказуючи на голослівне твердження або емоційний тригер).

Генеративна роль XLNet. Попередньо навчена модель XLNet теоретично здатна генерувати зв'язний текст на основі попередніх фраз діалогу. Це дає змогу використовувати її в ролі «опонента» користувача - співрозмовника, який висуває підступні аргументи. Генератор на основі XLNet формує маніпулятивні репліки, використовуючи приховані підміни у формулюваннях, емоційні маніпуляції або доводи, підлаштовані під профіль користувача. Іншими словами, модель адаптує створювані повідомлення під контекст і слабкі місця конкретного учня. Генерація здійснюється на основі великого корпусу фейкових новин, мережевих дискусій і маніпулятивних висловлювань, завдяки чому ШІ освоює риторику демагогії. Така модель здатна симулювати, наприклад, приховані директиви (коли наказ маскується під прохання або нешкідливу пораду) або двозначні пропагандистські гасла. С. І. Хаякава зазначав, що «*найнебезпечніші слова - ті, які здаються очевидними*» [6]. У політичній мові багато фраз мають нейтральний або доброзичливий вигляд, але несуть приховані приписи й аксіоми. Трансформер якраз дозволяє симулювати таку двозначність. Наприклад, заява «*ми просто захищаємо свої традиційні цінності*» на перший погляд цілком безневинна, проте залежно від контексту перетворюється на риторику винятковості та тиску на інакомислення. Тренажер, генеруючи подібні фрази, відтворює реалії маніпулятивної мови і тим самим розвиває навички критичного слухання та читання у користувача.

Комбінація генерації та аналізу. Одним із варіантів організації діалогу в ШІ-системах є **retrieval-based підхід**, тобто модель, яка не генерує відповідь «з нуля», а **обирає її з наперед сформованого набору реплік**, виходячи з релевантності до контексту діалогу. Такий підхід часто застосовується на практиці: дослідження показують, що моделі на основі BERT добре справляються з задачами вибору найбільш відповідної репліки з бази можливих відповідей [8]. Наприклад, можна створити базу типових маніпулятивних висловлювань (по кілька прикладів на кожен тип: апеляція до емоцій, хибна дихотомія, хибний авторитет тощо). Існують модифікації BERT для багатоваріантного діалогу, де модель приймає на вхід кілька останніх повідомлень і визначає найбільш релевантну відповідь [9]. У цьому випадку модель працює в режимі ранжування: на вхід подаються останні репліки діалогу та список кандидатів-відповідей з бази, а система оцінює релевантність кожної з них.

Наприклад, можна скласти базу типових маніпулятивних фраз/заголовків. Тоді BERT може працювати в режимі ранжування відповідей: на вхід подається поточний контекст діалогу і варіант відповіді з бази, модель оцінює їхню відповідність. Найкращий за оцінкою варіант вимовляється ШІ. Для нашого тренажера це означає, що ШІ-агент може зберігати бібліотеку стратегій (набір фраз для кожної маніпуляції, наприклад, тієї самої апеляції до емоцій або хибного вибору) та обирати з неї репліку залежно від розвитку розмови

Варто зазначити, що розподіл ролей «*BERT як аналітик - XLNet як провокатор*» дає змогу системі діяти на двох рівнях. З одного боку, одна модель провокує користувача на розбір складного висловлювання, граючи роль маніпулятора. Інша модель паралельно інтерпретує та коментує реакцію користувача. Завдяки цьому тренажер наближається до повноцінного інтелектуального співрозмовника, у якому кожен компонент - від вибору слів до стилю мовлення - підпорядкований навчальному завданню коментувати його реакцію. Таким чином, тренажер наближається до моделі повноцінного інтелектуального діалогу, в якому кожен елемент від вибору слів до стилю

підпорядкований навчальному завданню.

4. Bi-LSTM і механізм уваги в діалоговій системі

Рекурентні нейромережі, особливо їхні двоспрямовані модифікації (Bi-LSTM), і механізми уваги (attention) є потужними інструментами аналізу послідовного тексту. У контексті навчального діалогу їхня роль особливо важлива, оскільки вони забезпечують не тільки «розуміння» сказаного, а й контекстуальну чутливість до інтонацій, до логіки переходів, до імпліцитних зв'язків між фразами.

Модуль послідовності (Bi-LSTM). Двонаправлений LSTM обробляє послідовність реплік в обох напрямках, «запам'ятовуючи» як передісторію діалогу, так і подальший розвиток розмови. Це дає системі можливість відстежувати, як користувач будує аргументацію, до яких тем повертається, де допускає логічні збої або повторення. Наприклад, якщо співрозмовник схильний робити **стрибки від часткового до загального** (випадки хибного узагальнення), повертаючись до широких заяв без доказів, рекурентна мережа вловить цей патерн. У деяких роботах пропонувалося спочатку отримувати ембеддінги реплік через BERT, а потім пропускати їх через Bi-LSTM, щоб врахувати порядок реплік [12]. Такий підхід (застосований, наприклад, у моделі Speaker-Aware BERT) покращував якість розуміння діалогу. У нашому тренажері зв'язка «BERT + Bi-LSTM» дає змогу вибудовувати свого роду **когнітивний профіль** співрозмовника, розпізнаючи стійкі та вразливі ланки в його мовленнєвій поведінці. Так система зможе помітити, що, наприклад, користувач частіше приймає на віру статистичні твердження (можлива вразливість до апеляції до авторитету цифр) або, навпаки, гостро реагує на певні тригерні слова.

5. Модуль уваги (Attention)

Механізм **attention** дає змогу виокремити ключові слова та фрази, які мають найбільший вплив на інтерпретацію повідомлення. У рамках тренажера attention може відстежувати, на які частини висловлювання користувач звернув увагу, а які навпаки пропустив, хоча вони були семантично значущі. Наприклад, якщо користувач не помітив приховане застереження чи емоційно забарвлене слово, модель уваги це виявить. Після завершення репліки attention-субмодуль може підсвітити користувачеві фрагменти тексту, на яких сама модель сконцентрувалася як на значущих, у такий спосіб вказуючи на можливі хитрощі: «Зверни увагу, тут використане емоційно заряджене слово, щоб вплинути на тебе». Так реалізується механізм наочного зворотного зв'язку.

Хаякава наголошував, що реакція на слова часто випереджає усвідомлення їхнього сенсу: «іноді ми обурюємося ще до того, як зрозуміли, про що йдеться». Механізм attention якраз дає змогу «зловити» цей момент: з'ясувати, що саме в почутому викликало миттєву емоційну реакцію, і зробити цю інтуїтивну реактивність предметом аналізу. Візуалізація розподілу уваги допомагає культивувати метапізнання: після кожної сесії тренажер може надавати «теплову карту» слів, на які користувач реагував найемоційніше, забезпечуючи її поясненнями. Це своєрідне дзеркало, що допомагає користувачеві помітити не тільки сенс сказаного, а й власну реакцію на форму, інтонацію, контекст висловлювання. Раніше вже зазначалося, що поєднання **Attention+BiLSTM** показало високу точність при класифікації пропаганди [1]. У

нашому випадку ця зв'язка не тільки аналізує, а й спрямовує увагу користувача, підсвічуючи ключові повороти сенсу в діалозі. Це наближає форму спілкування до сократичного обговорення, де мета - не «перемогти» опонента, а зрозуміти, що саме викликає твою згоду чи незгоду.

6. Тематичне моделювання в навчальному діалозі

Дезінформація, як правило, не існує в ізоляції - вона вплетена в певні **наративні структури**. Це повторювані сюжети, які з часом посилюються і трансформуються, створюючи контекст, на якому одні аргументи здаються очевидними, а інші немислимими. Тематичне моделювання дає змогу «витягнути на світло» ці сюжети, показати, в яких координатах працює маніпуляція.

Найвідоміший класичний метод тематичного аналізу - LDA (латентне розміщення Діріхле [12]). У сучасній практиці застосовують і складніші алгоритми: BERTopic, NMF, тематичні моделі на основі трансформерів [11]. Їхнє завдання - виділити стійкі теми, кластери смислів, навколо яких будується комунікація. У контексті тренажера тематичне моделювання вирішує одразу кілька завдань.

По-перше, воно дає змогу диференціювати **мовні сценарії за тематикою**. Одна справа - маніпуляції у сфері охорони здоров'я (наприклад, навколо вакцинації), інша - у політичних гаслах, третя - в економічних дискусіях. Кожна тема потребує свого підходу, своїх прикладів, своєї стилістики мовленнєвих пасток. Тематичний аналіз зможе підлаштовувати зміст діалогу під ту сферу, яка близька користувачеві або в якій він частіше піддається впливу.

По-друге, тематичне моделювання дає змогу відстежувати тематичну варіативність реакцій користувача під час діалогу, накопичуючи контекстну інформацію про його типові зміщення в бік тієї чи іншої теми. Це дозволяє побічно аналізувати смислову стійкість користувача та його схильність до тематичних провокацій. Наприклад, під час однієї сесії співрозмовник зіткнувся з п'ятьма різними темами; у трьох із них він виявив схильність погоджуватися на емоційні аргументи, а у двох - чинив опір. Ці дані дають змогу проаналізувати смислову стійкість користувача: визначити, в яких галузях знань він найуразливіший до хитрощів. Хаякава застерігав від «стрибків рівня узагальнення» - ситуації, коли розмова про конкретний факт раптово йде в гаслове узагальнення: «це підло» → «суспільство вироджується» → «країна в небезпеці». Тематичне моделювання допомагає відкотити такий діалог назад - повернутися від абстракції до конкретики, поставити уточнювальне запитання: «а що саме в цій фразі нас тривожить?», «який прихований підтекст вона активує?».

По-третє, тематичні моделі можуть використовуватися для візуалізації «**карти тем**» діалогу. По завершенні тренування користувач може побачити схему тем, з якими він працював, і які слова активували ту чи іншу тему; які образи повторювалися; які гасла виникали найчастіше. Таким чином, користувач буквально дивиться в когнітивне дзеркало і спостерігає структуру свого сприйняття. Тематичне моделювання стає інструментом самоаналізу, що дає змогу виявити стереотипи та вразливості мислення. Це один зі шляхів до того, що Хаякава називав практикою «**смислової гігієни**» - свідомого контролю над мовою для профілактики маніпуляцій.

7. Ансамблеві методи в навчальному діалозі

Подібно до реального спілкування, де та сама людина може в різний час

використовувати різні стилі - логічні доводи, іронію, емоційний тиск, гру з фактами - в інтелектуальному тренажері доцільно застосовувати **ансамбль моделей**. Це дасть змогу створити багаторівневу систему, в якій різні алгоритми виконують різні ролі, імітуючи багатство живої мови.

Найочевидніший підхід - ансамбль із генеративної та дискримінативної моделей. Генератор (наприклад, XLNet або GPT) формує маніпулятивні репліки, імітуючи певну позицію в суперечці. Класифікатор (наприклад, BERT або Bi-LSTM) паралельно аналізує реакцію користувача: наскільки він згоден, які аргументи сприймає, а де сумнівається. Відбувається поділ праці: одна модель «грає», інша “оцінює”, що відповідає зв'язці «провокатор - наставник».

Більш просунутий ансамбль включає спеціалізовані детектори різних видів маніпуляцій. Наприклад, можна задіяти кілька вузькоспрямованих моделей, кожна з яких відстежує свій тип хитрощів: 1. **Логічні помилки** – модель аналізує аргументацію і ловить, наприклад, підміну тези або неправильне узагальнення. 2. **Емоційні пастки** – модель тонального аналізу виявляє приховані апеляції до емоцій і упереджено забарвлені вирази. 3. **Підміна аргументу** – детектор риторичних прийомів фіксує спроби звести складне питання до хибної дихотомії (чорно-білого вибору). 4. **Неправдива апеляція до авторитету** – модель відстежує заяви, які намагаються надати ваги словам за рахунок посилань на авторитет або статистику без достатніх підстав.

Завдяки такому ансамблю тренажер перетворюється на систему дидактичного аналізу, де кожен тип помилки відстежується і розбирається для користувача. Система може у зрозумілій формі повідомити результат: «*тут вас намагалися переконати за рахунок узагальнення*», «*у цьому місці використано підмінений аргумент*» тощо. Хаякава дуже влучно описував подібні випадки «мовного маскування», коли використання складної або псевдонаукової термінології створює видимість правди: «*люди часто вірять фразам не тому, що зрозуміли їх, а тому що вони звучать солідно*». Наш ансамбль якраз покликаний зривати це маскування, роблячи приховані прийоми явними.

Ансамбль моделей можна будувати не тільки за функціональним принципом, а й за поведінковим. Наприклад, один ШІ-агент у системі може грати роль скептика, інший - консерватора, третій - радикала-анархіста. Кожна з цих ролей використовує свою манеру мови і свої хитрощі, моделюючи свого роду «риторичні атаки» з різних позицій. Після діалогу система розбирає, який з аргументів виявився найскладнішим для користувача і чому. Таким чином, за рахунок ансамблевого підходу вибудовується система педагогічного розмаїття, де різні моделі кидають виклик по-різному, але узгоджено сприяють одній меті - навчитися помічати, розуміти і розбирати будь-які мовні маніпуляції.

8. Адаптація ШІ до стилю й аргументації користувача

Справжній діалог - це не просто обмін репліками, а й увага до особистості співрозмовника. Якщо ШІ не помічає як говорить людина, не вловлює її стиль, темп, особливості аргументації, він залишається всього лише механічним фільтром. Тому одне з ключових завдань тренажера - здатність системи підлаштовуватися під комунікативний профіль користувача. ШІ має реагувати на мовні звички, стиль мислення та очікування того, кого навчають. Одні користувачі схильні оперувати логікою, інші - емоціями; хтось викладає думки стисло та жорстко, хтось - витіювато й образно. Система розпізнає ці відмінності та враховує їх під час формування стратегії навчання.

Практично це означає наступне: якщо користувач схильний до декларативних тверджень - ШІ частіше ставитиме уточнювальні запитання; якщо користувач уникає конфлікту - тренажер м'яко спровокує його сформулювати свою позицію; а тільки-но учень звикає до одного типу маніпуляції, система поступово впроваджує інший, прийом, який раніше не зустрічався. Хаякава писав, що «людина часто не знає, якою мовою вона мислить - вона просто відтворює фрагменти риторики, засвоєної з оточення». Один із наведених ним прикладів - листи в газету, написані начебто людиною, яка говорить не від себе, а «голосом» газетної передовиці. Завдання тренажера - допомогти користувачеві почути свій власний голос і мислити самостійно, а не шаблонами.

Технічно адаптація реалізується через постійний аналіз дій користувача: профілювання його поведінки. Система групує типові реакції та висловлювання користувача, виділяючи профільні категорії (наприклад, схильність вірити статистиці або, навпаки, апелювати до особистого досвіду). - Тематичне профілювання. Відстежується, про що найчастіше говорить користувач, які теми для нього чутливі або цікаві. Це перегукується з тематичним моделюванням: ШІ визначає області, де учень вразливий до маніпуляцій. - Аналіз мовних шаблонів. Вивчається бажана граматика і структура мовлення користувача (довжина речень, частка оціночних суджень, спосіб побудови аргументів). Наприклад, велика кількість безапеляційних суджень сигналізує про звичку до категоричного мислення. - Відстеження емоційних реакцій. За непрямими ознаками - вибором відповідей, швидкістю і тоном реплік - модуль фіксує емоційний стан користувача (збентеження, обурення, згода тощо) у відповідь на ті чи інші хитрощі.

На базі цих даних система формує психолінгвістичний профіль і відповідно до нього змінює стиль власного мовлення. Десь тренажер додає іронію, десь посилює логічну зв'язність, а десь, навпаки, вмикає вкрадливу емоційну подачу - рівно настільки, щоб залишатися правдоподібним співрозмовником. Таке «тонке підстроювання» дає змогу вибудовувати довготривалу (а в ідеалі - нескінченну) програму риторичного вдосконалення користувача, не даючи йому застрягти на одному типі прийомів. Важливо, що з часом діалогові ситуації ускладнюються в міру зростання навичок - це підтримує ефект постійного розвитку і не дає тренуванню перетворитися на рутину.

Зрозуміло, подібна адаптивність вимагає серйозних обчислювальних ресурсів, особливо під час опрацювання мови в реальному часі. Щоб система залишалася інтерактивною і чуйною під час щоденного використання, в її архітектурі необхідно врахувати ресурсоефективність: використовувати оптимізовані моделі (наприклад, стислі версії BERT), виконувати найважчі обчислення на сервері або у фоновому режимі, а також застосовувати кешування результатів аналізу. Нижче описано архітектуру рішення, що враховує ці принципи.

9. Архітектура інтерактивного тренажера

Архітектура пропонованого тренажера мови будується за модульним принципом - кожен модуль відповідає за свою функцію, а спільно вони забезпечують гнучку і правдоподібну взаємодію з користувачем. Основні компоненти системи такі:

✳ Модуль генерації: відповідає за формування реплік ШІ. Він включає генеративну модель (наприклад, XLNet або іншу трансформерну мережу), здатну створювати реалістичні висловлювання із заданими маніпулятивними

властивостями. Цей модуль використовує бібліотеку сценаріїв і стратегій (хитрощів), а також знання про профіль користувача, щоб підібрати або згенерувати репліку, яка найбільше відповідає контексту діалогу. Генератор контролює тональність, лексику і складність фраз, намагаючись імітувати людину певного типу (агресивного сперечальника, «експерта»-авторитета тощо). За обмежених ресурсів модуль генерації може працювати в режимі retrieval (відбір заздалегідь підготовлених фраз за допомогою BERT, як описано вище) - це підвищує продуктивність, не жертвуючи реалістичністю діалогу.

✵ Модуль аналізу: виконує розбір відповідей користувача. До нього входить дискримінативна модель (наприклад, класифікатор на основі BERT або Bi-LSTM), яка оцінює кожне висловлювання користувача за кількома параметрами. Аналіз включає визначення ознак розпізнання маніпуляції (чи помітив користувач маніпулятивний прийом), емоційного стану (тональність відповіді) і ступеня згоди або сумніву. Модуль аналізу також відстежує використання користувачем певних ключових слів або логічних прийомів. Наприклад, якщо учень у відповідь починає сам вдаватися до надмірних узагальнень або цитування авторитетів, система це фіксує. Отримана інформація передається в модуль реакції та в модуль зворотного зв'язку для подальшого використання.

✵ Модуль реакції (управління діалогом): мозковий центр, який ухвалює рішення про подальший перебіг бесіди. Він отримує дані від модуля аналізу (про те, чи розпізнав користувач маніпуляцію, як відреагував емоційно, погодився чи ні) і обирає наступну стратегію. Наприклад, якщо аналіз показав, що користувач не розпізнав приховану директиву, модуль реакції може посилити цю саму маніпуляцію в наступній репліці або, навпаки, змінити тактику та спробувати інший прийом, щоб звернути увагу учня. Якщо ж користувач упевнено викрив маніпулятивний прийом, система може перейти до нової теми або підвищити рівень складності. Модуль реакції реалізує адаптивний сценарій: він планує діалог так, щоб поступово охопити всі цільові типи маніпуляцій (емоційні, логічні, семантичні) і водночас сфокусуватися на тих, що даються користувачеві найважче. По суті, тут закладається стратегія персонального навчання - аналог плану уроку, який гнучко перебудовується під час заняття.

✵ Модуль зворотного зв'язку: відповідає за пояснення і фіксацію пройденого матеріалу. Він накопичує ключові події діалогу: де користувач зробив помилку, де проявив проникливість, які маніпуляції зустрілися. Після (або під час) сесії модуль формує для користувача розгорнутий зворотний зв'язок. Це може бути текстовий розбір: наприклад, «На початку розмови ти не засумнівався в заяві з апеляцією до авторитету - зверни увагу: фраза містила посилання на невизначене дослідження, що типово для хибної апеляції». Зворотний зв'язок також презентується візуально, через підсвічування частин тексту або діаграми. Наприклад, можуть бути підсвічені всі коннотативно навантажені слова, які ШІ навмисно використовував, з поясненням їхньої прихованої оцінки; або показано хронологію логічних зрушень і узагальнень у суперечці. Модуль зворотного зв'язку в такий спосіб виконує роль тренера-коментатора: він не тільки вказує на помилки, а й рекомендує, як їх уникнути надалі (наприклад, радить при зустрічі з гучним твердженням шукати факти, тим самим боротися з підміною карти реальністю).

✵ Модуль адаптації: реалізує описані в попередньому розділі механізми підлаштування під користувача. Тут відбувається оновлення профілю

користувача - врахування його типових реакцій, прогресу та вподобань. Модуль адаптації отримує дані від модуля аналізу (мовні особливості, емоційні тригери) і від модуля зворотного зв'язку (засвоєні уроки, повторювані помилки). На основі цієї інформації він коригує параметри генерації та сценарій діалогу для наступних сесій. Наприклад, модуль адаптації може вирішити, що користувачеві настав час дати складніші завдання: додати багатозначних висловлювань або нових типів маніпуляцій, які раніше не зустрічалися. Також він може налаштувати стиль ведення діалогу - комусь підходить більш прямий, навіть жорсткий тон, а комусь - доброзичливий та ігровий. Усе це здійснюється автоматично, але на підставі чітких метрик: відсоток розпізнаних маніпулятивних прийомів, час реакції, частота емоційних сплесків тощо. У результаті кожна наступна сесія трохи не схожа на попередню, постійно кидаючи виклик саме тим навичкам, які потребують розвитку. Так досягається персоналізація навчання: двоє різних людей отримають згодом два різні досвіди взаємодії з системою, орієнтовані на їхні індивідуальні слабкі та сильні сторони.

✳ Модуль навчання (самонавчання системи): відповідає за накопичення знань і поліпшення самих моделей у міру експлуатації тренажера. Хоча основні моделі (генератори і класифікатори) навчені заздалегідь на великих корпусах, система здатна навчатися на даних, що збираються в діалогах з користувачами. Модуль навчання періодично аналізує журнал діалогів (зі знеособленням даних) у пошуках нових патернів: наприклад, нових видів маніпуляцій, на які користувачі реагують помилково, або несподіваних відповідей, не передбачених сценарієм. На основі цього він може ініціювати донавчання генеративної моделі (щоб та краще імітувала сучасні тактики маніпуляторів) або оновлення класифікаторів (щоб врахувати більше варіантів відповідей користувача). Крім того, модуль навчання займається оновленням бази знань: додає в бібліотеку нові приклади прийомів, нові теми, актуальні інформаційні приводи. Це дає змогу системі не застарівати і крокувати в ногу з інформаційним простором, що розвивається. В архітектурі закладено принцип continual learning - безперервного поліпшення без необхідності вручну перенавчати всю систему з нуля. За рахунок цього тренажер з часом стає тільки ефективнішим, спираючись на широкий досвід взаємодії з різними користувачами.

Архітектурні принципи та інтерфейс. Запропонована модульна архітектура забезпечує гнучкість і прозорість роботи системи. Поділ на генерацію та аналіз, з одного боку, дає змогу використовувати ансамбль вузькоспеціалізованих моделей (як описано вище), а з іншого - робить поведінку ШІ інтерпретованою. Кожному компоненту відповідають зрозумілі функції, і помилки можна локалізувати (наприклад, якщо некоректну фразу було згенеровано, її можна вилучити з бази; якщо класифікатор невірно зрозумів користувача - скорегувати його навчання). Для підвищення ресурсоефективності критично важливими є оптимізації: важкі мовні моделі можна запускати на серверному боці або використовувати їх зі зниженою точністю за умови обмежених ресурсів, а локально на пристрої користувача виконувати лише найважливіші обчислення (наприклад, швидкий аналіз наміру). У разі відсутності з'єднання тренажер здатний працювати у спрощеному режимі, покладаючись на заздалегідь завантажений сценарій і локальні моделі - це уможливить щоденне використання, яке не потребує спеціальних умов.

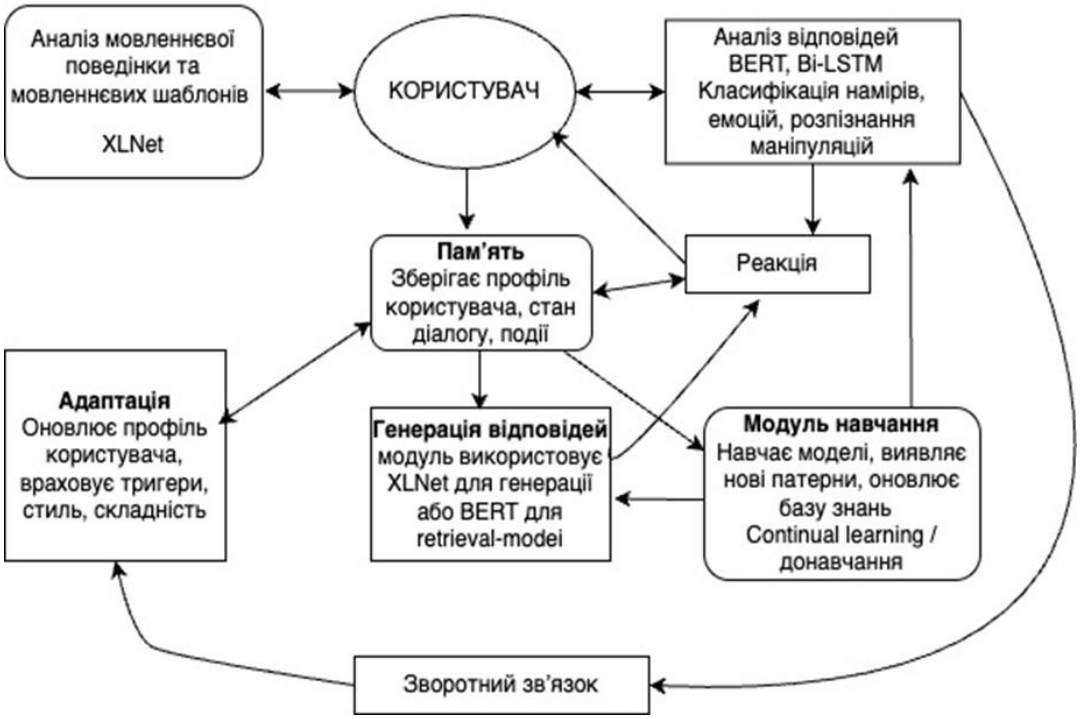

Рис. 1. Архітектура діалогового тренажера з адаптивним навчанням та зворотним зв'язком

Інтерфейс тренажера спроектований як звичний чат-діалог: користувач спілкується з віртуальним співрозмовником через текст (у перспективі - голосом), ставить йому запитання або відповідає на провокації. У процесі діалогу інтерфейс надає візуальні підказки: наприклад, неявно виділяє слова, на які варто було б звернути увагу, або пропонує натиснути кнопку «Чому це важливо?», якщо користувач розгублений. Після завершення кожної сесії користувачеві відображається інтерактивний звіт: виділені фрагменти його листування з колірним підсвічуванням (наприклад, червоним – місця, де прихована емоційна маніпуляція, жовтим - можливі логічні маніпулятивні прийоми), пояснення від системи та поради. Також показуються динамічні показники: скільки маніпуляцій він успішно розпізнав, над чим прогрес, які нові типи маніпулятивних прийомів зустрілися сьогодні. Такий дружній інтерфейс перетворює навчання на гру-викриття. За рахунок щоденних сесій (навіть коротких, 5-10 хвилин) користувач поступово звикає аналізувати інформацію автоматично - критичне мислення стає частиною його рутини. В ідеалі використання системи стане звичкою і стане настільки ж буденною і необхідною справою, як ранкова зарядка чи чищення зубів - тренажер щодня тренуватиме «розумову поставу» користувача. Завдяки інтерактивності та чуйності ШІ-співрозмовника кожен день спілкування з системою змістовний і трохи відрізняється від попереднього, що підтримує мотивацію та інтерес. Ми переконані, що така інтеграція ШІ та ідей Хаякави про мовну усвідомленість дасть змогу ефективно протистояти маніпуляціям і виховати в широкої аудиторії навичку здорового скепсису до слів.

Висновки

У результаті проведеного дослідження було запропоновано нову концепцію використання сучасних моделей обробки природної мови (зокрема BERT, XLNet, Bi-LSTM) у структурі інтелектуального діалогового тренажера,

метою якого є розвиток критичного мислення користувачів. Відмінність підходу полягає в акценті на навчанні людини через інтерактивну взаємодію з ШІ, що не лише імітує реальні комунікативні ситуації, але й адаптує власні стратегії мовної поведінки на основі психолінгвістичного профілю користувача.

Запропонована архітектура тренажера дозволяє реалізовувати сценарії мовленнєвих маніпуляцій (апеляція до емоцій, хибні узагальнення, авторитарні апеляції тощо), що ґрунтуються на класифікації, описаній С. І. Хаякавою. Було доведено доцільність адаптації математичних методів виявлення дезінформації до навчального контексту з метою формування навичок розпізнавання маніпуляцій.

У статті також окреслено принципи проєктування адаптивної гібридної архітектури тренажера, включаючи підсистеми сценарного керування, мовної генерації, адаптації, профілювання, зворотного зв’язку та тематичного модулювання. Обґрунтовано необхідність розробки ефективної онтологічної моделі для забезпечення семантичної узгодженості всіх компонентів.

Список літератури

1. Падалко, Г. А. Моделі, методи та інформаційні технології виявлення та аналізу текстової дезінформації та пропаганди у соціальних мережах : дис. ... д-ра філос. наук : 122 / Падалко Гліб Анатолійович. – Харків, 2025. – 320 с.
2. Khan, J. Y. [та ін.]. A Benchmark Study of Machine Learning Models for Online Fake News Detection // Machine Learning with Applications. 2021. Vol. 4.
3. Kumar, R., Goswami, A., Narang, P. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach // Multimedia Tools and Applications. 2021. Vol. 80 (8). P. 11765–11788.
4. Padalko, H., Chomko, V., Chumachenko, D. A novel approach to fake news classification using LSTM-based deep learning models // Frontiers in Big Data. 2024. Vol. 6.
5. McLuhan, M. Understanding Media: The Extensions of Man. New York : McGraw-Hill, 1964.
6. Hayakawa, S. I. Language in Thought and Action. New York : Harcourt Brace Jovanovich, 1978.
7. Socratic AI Against Disinformation: Improving Critical Thinking to Recognize Disinformation Using Socratic AI. ResearchGate. URL: https://www.researchgate.net/publication/371710581 (дата звернення: 30.05.2025).
8. Gu, X., Wang, T., Liu, X. Fine-grained Post-training for Improving Retrieval-based Dialogue Systems // Proceedings of the NAACL-HLT. 2021.
9. Gu, X., Zhang, H., Dai, X. et al. Speaker-Aware BERT for Multi-Turn Response Selection in Dialogue Systems // arXiv preprint. 2020. arXiv:2002.05897.
10. Cevher, M., Göktepe, H., Özyer, T. BERT-based Response Selection in Dialogue Systems Using Utterance Attention Mechanisms // ResearchGate. 2022.
11. Gu, J., Cai, D., Zhang, Y., Wang, H. Speaker-Aware BERT for Multi-Turn Response Selection in Dialog System [Електронний ресурс]. – Режим доступу: https://arxiv.org/abs/2002.05897 (дата звернення: 30.05.2025).
12. Blei, D. M., Ng, A. Y., Jordan, M. I. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. Vol. 3 (Jan). P. 993–1022.
13. Grootendorst M. BERTopic: Neural topic modeling with class-based TF-IDF and embeddings [Електронний ресурс]. – Режим доступу: https://maartengr.github.io/BERTopic (дата звернення: 30.05.2025).

References

1. Padalko, H. A. Models, Methods and Information Technologies for Detecting and Analyzing Textual Disinformation and Propaganda in Social Networks [PhD thesis, Kharkiv Aviation Institute]. Kharkiv, 2025.
2. Khan, J. Y., Islam, M. T. K., Afroz, S., Uddin, G., & Iqbal, A. (2021). A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications, 4, 100033.
3. Kumar, R., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, 80(8), 11765–11788.
4. Padalko, H., Chomko, V., & Chumachenko, D. (2024). A novel approach to fake news classification using LSTM-based deep learning models. Frontiers in Big Data, 6. https://doi.org/10.3389/fdata.2023.123456
5. McLuhan, M. (1964). Understanding media: The extensions of man. New York: McGraw-Hill.
6. Hayakawa, S. I. (1978). Language in thought and action. New York: Harcourt Brace Jovanovich.
7. Socratic AI Against Disinformation. (n.d.). Improving critical thinking to recognize disinformation using Socratic AI. ResearchGate. Retrieved May 30, 2025, from https://www.researchgate.net/publication/371710581
8. Gu, X., Wang, T., & Liu, X. (2021). Fine-grained post-training for improving retrieval-based dialogue systems. In Proceedings of the NAACL-HLT.
9. Gu, X., Zhang, H., Dai, X., et al. (2020). Speaker-aware BERT for multi-turn response selection in dialogue systems. arXiv preprint, arXiv:2002.05897.
10. Cevher, M., Göktepe, H., & Özyer, T. (2022). BERT-based response selection in dialogue systems using utterance attention mechanisms. ResearchGate.
11. Gu, J., Cai, D., Zhang, Y., & Wang, H. (2020). Speaker-Aware BERT for Multi-Turn Response Selection in Dialog System. arXiv. Retrieved May 30, 2025, from https://arxiv.org/abs/2002.05897
12. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.
13. Grootendorst, M. (2022). BERTopic: Neural topic modeling with class-based TF-IDF and embeddings. Retrieved May 30, 2025, from https://maartengr.github.io/BERTopic

Надійшла до редакції 13.05.2025, розглянута на редколегії 13.05.2025

An interactive model for teaching critical thinking through AI dialogue

The article presents an interdisciplinary study devoted to the development of an interactive dialogue simulator that uses artificial intelligence to develop critical thinking skills in users. The central idea of the work is a paradigm shift: it is not humans who teach AI, but AI, adapting to the user, acts as a coach and interlocutor, modeling various types of manipulative speech behavior. The simulator takes the form of a daily dialogue that imitates real communication situations, with the aim of gradually developing the user's ability to recognize disinformation, hidden directives, logical errors, and psychological influences.

The concept of the simulator is inspired by the ideas of S. I. Hayakawa, set out in his fundamental work “Language in Action” [6], which analyzes typical linguistic manipulations that reduce the quality of thinking. It is this conceptual apparatus that became the basis for the simulator's scenario module, where each type of influence — such as appeals to emotions, authority, or false dichotomies — is reproduced using appropriate speech generation algorithms.

The technological part of the study analyzes the possibilities of using modern language models (BERT, XLNet, Bi-LSTM) in the context of educational dialogue interaction. In particular, it considers how disinformation classification models can be adapted to the learning process, taking into account the user's psycholinguistic profile. A hybrid system architecture is proposed with modules for adaptation, thematic modeling, feedback, semantic analysis, and support for long-term interaction.

The article aims to create a basis for the development of a full-fledged digital tool — a simulator capable not only of detecting disinformation but also of forming a culture of conscious speech and perception of information. The proposed system can be integrated into educational platforms, in particular for individualized training of students, journalists, teachers, and other professions that require critical analysis of information. This approach contributes not only to the development of cognitive skills but also to the formation of resistance to information influences in the digital age.

Key words: critical thinking; artificial intelligence; disinformation; language manipulation; BERT; XLNet; simulator; dialogue system; Hayakawa; educational technologies.

Відомості про авторів:

Чухрай Андрій Григорович – проф., доктор техн. наук, проф. каф. інженерії програмного забезпечення Національного аерокосм. університету «ХАІ», Харків, Україна, a.chukhray@khai.edu , ORCID 0000-0002-8075-3664.

Шаповал Микита Володимрович – аспірант каф. інженерії програмного забезпечення Національного аерокосмічного університету «ХАІ», Харків, Україна, m.v.shapoval@khai.edu , ORCID 0009-0009-6899-9539.

Мандрикова Людмила Василівна – доцент, канд. техн. наук, доцент каф. інженерії програмного забезпечення Національного аерокосм. університету «ХАІ», Харків, Україна, l.mandrikova@khai.edu, ORCID 0000-0003-1781-1662.

About the authors:

Chukhray Andrey – professor, doctor of technical sciences, professor of the department of software engineering, faculty of software engineering and business of the National Aerospace University "KhAI", Kharkiv, Ukraine, e-mail: a.chukhray@khai.edu , ORCID 0000-0002-8075-3664.

Shapoval Mykyta – PhD student at the department of software engineering, National Aerospace University "KhAI ", Kharkiv, Ukraine, email: m.v.shapoval@khai.edu , ORCID 0009-0009-6899-9539.

Ludmila Mandrikova – associate professor, candidate of technical sciences, associate professor of the department of Software Engineering of the National Aerospace University "KhAI ", Kharkiv, Ukraine, e-mail: l.mandrikova@khai.edu , ORCID 0000-0003-1781-1662