УДК 81`322: 81`42: 811.612.91

DOI: 10.24025/2707-0573.12.2025.345012

Вікторія Мусійчук

ПІЛОТНИЙ ЕКСПЕРИМЕНТ З АВТОМАТИЧНОГО ВИЯВЛЕННЯ В'ЄТНАМСЬКИХ МЕДІАНАРАТИВІВ ПРО РОСІЙСЬКО-УКРАЇНСЬКУ ВІЙНУ ПРИ ОБМЕЖЕНИХ РЕСУРСАХ

У статті описано проведення експерименту з автоматизованого виділення наративів про російсько-українську війну з в'єтнамських медіатекстів. Традиційний наративний аналіз поєднано з цифровими методами гуманітарних наук. Дослідження базується на абдуктивному підході, що поєднує двосторонню обробку даних. Послідовні етапи експерименту передбачають використання сучасних інструментів обробки природної мови, адаптованих спеціально для в'єтнамської мови та в умовах обмежених ресурсів.

Ключові слова: *в'єтнамська мова, медіа, наративи, PhoBERT, кластеризація, російсько-українська війна.*

1. Вступ

Наративи у медіатекстах відіграють ключову роль у формуванні громадської думки щодо міжнародних конфліктів, зокрема російсько-української війни. У в'єтнамському медіапросторі ці наративи характеризуються поляризацією та неоднозначністю. Державні ЗМІ, декларуючи нейтральну позицію В'єтнаму, водночас послуговуються проросійськими інтерпретаціями подій та термінології. Соціальні мережі наповнені як проросійськими, так і проукраїнськими настроями. Тому системне вивчення природи формування в'єтнамських медіанаративів є надзвичайно актуальним як з лінгвістичної точки зору, так і для подальшого застосування результатів цих досліджень для вироблення ефективних стратегій комунікації на міжнародному дипломатичному рівні.

Дослідження в'єтнамських медіанаративів російсько-української війни передбачає створення корпусу новин, повідомлень, статей, дописів з сайтів ЗМІ, соцмереж, блогерських платформ. З 2022 року донині (а для якісного дослідження варто брати період з 2014 року) – це має бути величезний обсяг даних. Сучасні наративні дослідження для опрацювання великих масивів даних передбачають використання методів цифрової гуманітаристики та штучного інтелекту, а також поєднання якісних підходів з кількісними. Такі

дослідження мають загальні обмеження, зокрема, це застосування ресурсомістких, як технічно, так і фінансово, технологій. Основою для дослідження наративів є комплексний підхід до збору і опрацювання корпусу текстів, зокрема відбір релевантних джерел із очищенням та структуризацією даних. Необхідно зазначити, що для в'єтнамської мови уже на цьому етапі постає чимало викликів: відсутність анотованих датасетів; обмежена кількість як загальномовних, так і тематичних медійних корпусів; обмежений доступ до багаторівневих даних на в'єтнамських медійних сайтах; окрім звичайного очищення даних, в'єтнамські тексти вимагають додаткової токенізації для виділення слів, адже пробіли в тексті стоять не між словами, а між морфемами; більшість ШІ-моделей, NLP-інструментів навчені на англомовних корпусах, і їхня точність для в'єтнамської мови є істотно нижчою; розробка NLP-інструментів для в'єтнамської мови ведеться не так активно, а попередні розробки дуже швидко застарівають та не можуть бути застосовані через несумісність з сучасними системами. Така обмеженість ресурсів ускладнює масштабні дослідження. Наступний крок наративного аналізу – власне виокремлення та інтерпретація наративів, що базується на аналізі текстового корпусу – також потребує особливого підходу з урахуванням особливостей для в'єтнамського мовного матеріалу. Отже, завданням цієї статті є пристосування наявних методів та інструментів автоматичного аналізу в'єтнамськомовних текстів з метою виробити алгоритм для виявлення наративів в умовах обмежених ресурсів.

2. Теоретичне підґрунтя

Наративний аналіз об'єднує структуралістські, критично-дискурсивні підходи, фрейм-аналіз та комп'ютерний аналіз. Спадщина класичної наратології бере свій початок у структуралізмі, зокрема у роботах В. Проппа, який виокремив 31 функцію казки, що стали основою для виявлення універсальних наративних паттернів (Propp, 1968). Р. Барт розвинув цю модель, запропонувавши трирівневу систему аналізу: рівень функцій, рівень дій та рівень наративу, де кожен рівень інтегрує попередній у складнішу семантичну структуру (Barthes, 1975). А. Греймас переосмислив проппівські персонажі в актантну модель (суб'єкт-об'єкт, відправник-отримувач, помічник-опонент), яка дозволяє аналізувати глибинні семантичні структури незалежно від поверхневих лінгвістичних форм (Greimas, 1983).

Критичний дискурс-аналіз, представлений у роботах Н. Ферклафа та Т. ван Дейка, трактує наратив як соціальну практику, що відтворює та легітимізує владні відносини через мовні вибори (Fairclough, 1995; van Dijk, 1998). Ферклаф запропонував трикомпонентну модель аналізу: текст (опис лінгвістичних властивостей), дискурсивна практика (інтерпретація процесів виробництва/споживання) та соціальна практика (пояснення ідеологічних наслідків) (Fairclough, 1989, 1995). Ван Дейк додав соціо-когнітивний вимір, акцентуючи увагу на ментальних моделях та ідеологічних схемах (van Dijk, 2008).

Теорія фреймінгу (Entman, 1993; Goffman, 1974) пояснює, як медіа подають події в обмежених рамках (фреймах), вибірково акцентуючи певні аспекти реальності: визначають проблему, діагностують причини, виносять моральні судження та пропонують рішення (Entman, 1993). Комп'ютерний наративний аналіз інтегрує зазначені вище теорії з NLP-методами (Piper, 2021). Методи на основі графів, як-от наративні мапи (Narrative Maps) та наративні стежки (Narrative Trails), використовують семантичні векторні представлення

текстів, щоб вибудовувати послідовні й цілісні сюжетні лінії, оптимізуючи їхню змістову узгодженість (Keith, 2020; German, 2025). Водночас подієцентричні підходи (event-centric frameworks) зосереджуються на виокремленні подій, дійових осіб та їхніх взаємозв'язків, що дає змогу виявляти характер фреймування цих подій у медійному дискурсі (Das, 2024; Levi, 2020).

Для нашого завдання автоматизованого виявлення наративів доцільним видається подієцентричний підхід, заснований на виділенні ключових слів, з елементами фреймування та кластеризації на основі ембедінгу (семантичних векторів).

3. Огляд літератури

Дослідження сприйняття російсько-української війни у в'єтнамському медіапросторі є нечисленними. Аналіз понад двохсот дописів в'єтнамського сегменту мережі Facebook протягом першого року конфлікту, проведений Май Тхі Данг Тху, виявив вісім ключових наративних фреймів, серед яких домінували антизахідні (США та Європа), антиукраїнські та проросійські настрої, що відображають національні настрої та геополітичні чинники. Підкреслено критичну роль соціальних медіа як впливового інструменту для просування геополітичних наративів і формування громадської думки у В'єтнамі. Дослідниця використала ручний контент-аналіз для виявлення прихованих конотацій та організації їх у фрейми, а також кількісні методи, як-от частотний аналіз, дослідження динаміки та кореляції між фреймами (Mai, 2025).

У двох працях в'єтнамських дослідників використано метод дискурс-аналізу для виявлення особливостей проросійських наративів у в'єтнамському кіберпросторі після початку повномасштабного вторгнення Росії в Україну. Аналіз 28 активних Facebook-груп засвідчив їхню функцію «ехокамер», що систематично посилюють дискурс, який легітимізує російську агресію, позитивно конотуючи москву та негативно репрезентуючи Україну й Захід. Дослідники пов'язують цю тенденцію з комплексом факторів, зокрема історичною ностальгією за радянським періодом та стійкими антизахідними сентиментами. Такі онлайн-наративи сприяють формуванню пропагандистського середовища, яке дозволяє в'єтнамському уряду, зокрема його консервативному крилу, утримувати нейтральну позицію у публічному дискурсі, унеможливлюючи домінування проукраїнських інтерпретацій (Hoang, 2022, 2025).

То Мінь Шон дослідив полярність громадської думки щодо конфлікту між Росією та Україною, зважаючи на в'єтнамську офіційну політику «принципового нейтралітету». Ця позиція є дипломатично складною через міцне партнерство з Росією та історичну ностальгію за радянською допомогою В'єтнаму, що й призвело, на думку дослідника, до внутрішнього розколу громадської думки. Хоча частина населення висловлює пропутінські настрої, інші занепокоєні питанням національного суверенітету та міжнародного права (To, 2022).

Загалом, наявні дослідження переважно сфокусовані на аналізі соцмереж з використанням якісних методів дослідження, як-от контент-аналіз та дискурс-аналіз, що є виправданим на невеликих вибірках даних. Однак для великих масивів даних не обійтися без методів комп'ютерної обробки та аналізу. Отже, метою цієї статті є тестування методології автоматичного виявлення в'єтнамських медіанаративів про російсько-українську війну за

умов обмежених ресурсів. Для пілотного експерименту зібрано невеликий корпус з новинного сайту «В'єтнамської інформаційної агенції» (Báo tin tức): 80 новин за період січень-квітень 2022 року та 80 новин за період січень-квітень 2025 року, відібраних за пошуковим запитом «Україна».

4. Методи та матеріал дослідження

Дослідження спроєктовано як пілотний експеримент для тестування методології автоматичного видобування медіанаративів при обмежених ресурсах. Корпус складається зі 160 в'єтнамських новинних текстів з сайту «В'єтнамської інформаційної агенції», вибраних за пошуковим словом «Україна» у кількох варіаціях його написання в'єтнамською мовою, а саме *Ukraine*, *Ukraina*, *Ucraina*. Усі новини про російсько-українську війну та інші новини про Україну було розмежовано, про що йтиметься нижче. Вибір джерела для корпусу зумовлений тим, що «В'єтнамська інформагенція» є офіційним державним інформаційним органом, який є джерелом новин для багатьох інших в'єтнамських ЗМІ. Вибір періоду чотири місяці 2022 року та чотири місяці 2025 року дозволив забезпечити на невеликому обсязі даних тематичне та подієве розмаїття (настрої перед війною, масова евакуація та міграція, окупація, воєнні дії, обстріли тилових населених пунктів, дипломатичні перемовини тощо). Вибір розміру корпусу визначений метою швидкого тестування, можливості ручної перевірки та практичними обмеженнями (приблизно 12 гігабайтів оперативної пам'яті, максимальна тривалість безперервної роботи 12 годин, та поточні обмеження на пропускну спроможність при завантаженні/вивантаженні даних з Google Drive) безкоштовної версії хмарної платформи для розробки на мові Python – Google Colab (Google Colaboratory), що використовувалася для автоматичної обробки.

Корпус було зібрано методом вебскрейпінгу з використанням платформи ParseHub (ParseHub). Ця платформа дозволяє збирати тексти, одночасно очищуючи від зайвих артефактів, що полегшує подальшу роботу з корпусом.

Базовий препроцесинг було здійснено у середовищі Google Colab за допомогою бібліотеки Underthesea-Vietnamese NLP Toolkit (Underthesea, 2025). Інструмент виконує токенізацію у вигляді об'єднання морфем у слова. Цей крок є критично важливим для подальшого аналізу, адже в'єтнамська мова не має пробілів між словами на письмі, натомість має пробіли між морфемами. Таким чином, проведення такої токенізації є обов'язковим для подальшої можливості виділення ключових слів, тематичної рубрикації та інших необхідних для наративного дослідження функцій. З іншого боку, в'єтнамський текст не потребує лематизації, адже в'єтнамська мова не має словозміни.

Виділення наративів відбувалося на основі подієцентричного підходу, який спирається на ключові слова для подій, персонажів та тем. Також було здійснено групування виділених фрагментів та відповідне фреймування й кластеризацію для подальшої можливості інтерпретації наративів.

Загальний хід дослідження ґрунтується на абдуктивному підході, запропонованому Ч. Пірсом, який поєднує переваги індукції та дедукції (Burch, 2024). У цьому дослідженні це реалізовано двома взаємодоповнювальними напрямами:

1) індуктивний: ключові слова → теми текстів → великі тематичні групи;

2) дедуктивний: початкова кластеризація → конкретні прояви тем.

Для автоматичного видобування ключових слів і фраз із текстів використано KeyBERT, який дозволяє це робити без попереднього навчання (Grootendorst, 2020). На відміну від традиційних статистичних методів, які спираються на частотність слів, KeyBERT використовує семантичні ембедінги (числові представлення слів у багатовимірному просторі), які захоплюють змістовне значення слів. Алгоритм KeyBERT працює в кілька етапів. По-перше, весь документ кодується в один вектор (багатовимірну точку) за допомогою моделі Sentence-BERT, що представляє загальний зміст документа. Далі вилучаються всі окремі слова та короткі фрази, й кожна із них також кодується у вектор. Алгоритм порівнює кожного кандидата із вектором цілого документа, обчислюючи їхню семантичну подібність. Найбільш подібні до документа слова та фрази обираються як ключові. Основною перевагою KeyBERT є те, що він не потребує попереднього навчання на спеціалізованих корпусах та підтримує будь-яку мову, для якої є модель BERT, що робить його особливо корисним для низькоресурсних мов, як-от в'єтнамська.

У нашому дослідженні для цієї мети використано модель VoVanPhuc/sup-SimCSE-VietNamese-phobert-base – це передова модель штучного інтелекту, спеціалізована на створенні векторних представлень речень для в'єтнамської мови (VoVanPhuc, 2024). Вона базується на моделі PhoBERT, яку було натреновано на в'єтнамських текстах, що забезпечує їй розуміння як повсякденної, так і професійної лексики (Nguyen, 2020). Крім того, сентимент-аналіз з PhoBERT показав точність 93.12% у класифікації новин на позитивні / негативні / нейтральні (Nguyen, 2021), демонструючи потенціал моделі для аналізу наративів. PhoBERT перетворює кожен текст на числовий вектор розмірністю 768, що дозволяє порівнювати тексти математичним шляхом: тексти зі схожим змістом матимуть числові представлення, близькі одне до одного у цьому просторі. SimCSE-Vietnamese – це окрема модель, оптимізована для порівняння семантичної подібності текстів у в'єтнамській мові. Вона натренована на множині пар текстів, де парафрази позначені як подібні, а семантично не пов'язані тексти – як неподібні. Це навчання робить модель особливо чутливою до справжніх змістовних схожостей. Використання PhoBERT+SimCSE забезпечило дві ключові переваги для цього дослідження: семантичні представлення, адаптовані до в'єтнамської мови, і можливість швидкої обробки в Google Colab.

Для кластеризації текстів у цьому експерименті використано два комплементарні методи: K-means та HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). K-means – це алгоритм, що поділяє тексти на певну кількість груп (кластерів). Алгоритм працює ітеративно: спочатку він випадково вибирає кілька «центрів» (чисельні точки в семантичному просторі, які представляють центр кожної групи), до яких зараховуються тексти. Після цього алгоритм перераховує позиції центрів на основі всіх текстів, зарахованих до кожної групи, й перерозподіляє групи. Процес продовжується, поки центри не перестануть суттєво змінюватися, що свідчить про стабільність групування (Abiodun, 2023).

У цьому дослідженні метод кластеризації HDBSCAN застосовано як експериментальний інструмент для групування в'єтнамських текстів про російсько-українську війну за їхньою семантичною подібністю, використовуючи ембедінги PhoBERT/SimCSE як вхідні дані. HDBSCAN реалізує кластеризацію на основі щільності з автоматичним виявленням

викидів, а PhoBERT-ембедінги забезпечують семантичну якість (Nguyen et al., 2023). HDBSCAN належить до щільнісних методів кластеризації: він шукає в багатовимірному просторі ділянки, де тексти розташовані «густо» (мають подібні значення ембедінгів), і відокремлює їх від зон низької щільності, які інтерпретуються як шум або маргінальні випадки (McInnes, 2017). На відміну від K-means, HDBSCAN не вимагає наперед задавати кількість кластерів, а натомість має параметр мінімального розміру кластера, що краще відповідає ситуації, коли невідомо, скільки саме різних наративних груп є в корпусі.

GPT-4 (ChatGPT) було використано як допоміжний інструмент для розподілу ключових фраз на події, персонажі та теми з подальшою ручною верифікацією, а також для якісної анотації попередньо виділених кластерів.

Дослідження організовано як паралельна послідовність двох напрямів обробки. Для обох напрямів спочатку створено корпус в'єтнамських медіатекстів та здійснено препроцесинг вихідних даних та ембедінг через PhoBERT+SimCSE. Далі у пешому напрямі йде видобування ключових фраз за допомогою KeyBERT та групування з GPT-4. У другому напрямі – кластеризація через K-means і HDBSCAN та інтерпретація за допомогою GPT-4.

Схема 1. Алгоритм автоматичного виявлення наративів у в'єтнамських медіатекстах з абдуктивним підходом

5. Результати дослідження

5.1. Загальні характеристики корпусу та розподіл даних

Для дослідження було створено експериментальний мінікорпус із 160 в'єтнамських новинних текстів, видобутих за допомогою платформи ParseHub. Корпус було збережено у форматі .csv з усіма вихідними даними (як-от дата, лінк, автор, заголовок), де кожна новина знаходиться в окремому рядку. Такий формат дозволяє в одну операцію обробляти багато текстів, водночас не зливаючи усі тексти в один. Попередню обробку тексту було проведено за допомогою інструменту Underthesea, а саме його частини word_tokenizer, для розбиття тексту на слова. Проблема виділення меж слова у в'єтнамській мові така складна й неоднозначна, що жоден автоматичний інструмент досі не може вирішити це завдання без похибок. Зокрема, на

нашому матеріалі серед іншого було помічено, що після обробки в одне слово об'єднувалися іноземні імена та прізвища або власні назви з кількох слів (наприклад, *Donald_Trump*, *Eurovision_News*), що для нашого дослідження було не вадою, а перевагою.

5.2. Від ключових слів до тематичних груп

У першій частині дослідження кожен текст було піддано структурованому аналізу за допомогою методу KeyBERT, що функціонує на основі моделі PhoBERT SimCSE-Vietnamese. Цей відбір інструментарію було обґрунтовано тим, що SimCSE-Vietnamese є передовою моделлю, спеціалізованою на створенні контекстуально чутливих векторних представлень речень для в'єтнамської мови з використанням методу Simple Contrastive Learning of Sentence Embeddings на основі архітектури PhoBERT. На цьому етапі з корпусу текстів було автоматично виділено всі можливі n-грами довжиною від 1 до 3 слів, після чого відібрано найчастотніші з них разом із відповідними векторними представленнями. Це дозволило встановити базові семантичні одиниці наративного дискурсу. Було проекспериментовано видобувати 50, 100 та 200 ключових фраз. Необхідно зазначити, що збільшення кількості ключових слів мало тенденції до повторів. Наприклад, один персонаж повторювався кілька разів у сполученні з різними діями або одна й та сама подія подавалася у різних граматичних вираженнях:

tham_gia của mỹ *0.4738* 'участь Америки',
mỹ đã tham_gia *0.4742* 'Америка брала участь',
kirill_dmitriev đưa ra *0.4584* 'Кіріл Дмітрієв запропонував',
dmitriev cho biết *0.4677* 'Дмітрієв сказав',
kirill_dmitriev báo_cáo *0.4756* 'Кіріл Дмітрієв повідомив'.

Крім того, на початковому етапі виділення ключових фраз відбувалось без використання стоп-слів, тому до списків потрапили такі нерелевантні сполуки:

viện_trợ quân_sự đã *0.5660* 'військова допомога вже' – *đã* є граматичним показником минулого часу та не має значення без дієслова,

hỗ_trợ quân_sự hơn *0.5643* 'військова підтримка більше ніж' – *hơn* є граматичним показником ступеню порівняння,

chuyên_gia quân_sự và *0.5087* 'військові експерти та' – *và* є сполучником.

Отже, використання списку стоп-слів для цього етапу є вкрай важливим. При чому, окрім звичайних для таких списків службових слів, імовірно варто додати й повнозначні слова, які часто трапляються у новинах, але не завжди мають вагоме значення для виділення ключових фраз, як-от: *сказав*, *повідомив* тощо.

Ще одним недоліком цієї фази експерименту є те, що необхідно заздалегідь визначити кількість ключових слів і фраз. Для великих масивів даних це може бути незручно.

Наступним завданням було перетворити ключові фрази з KeyBERT на наративні мітки, зокрема події й персонажі. Цей підхід дав можливість згрупувати схожі ключові слова і фрази, спростити їх та систематизувати наративи на двох рівнях абстракції. Для виконання завдання було використано ChatGPT. У результаті було створено списки слів на позначення подій та персонажів. Наприклад, ключові слова на позначення подій: *tấn công – атака, pháo kích – обстріл, xâm lược – вторгнення, di tản – евакуація, viện trợ nhân*

đạo – гуманітарна допомога, đàm phán – переговори, chiến sự – військові дії. Серед персонажів, або акторів, зокрема було виділено такі: *Ukraine – Україна, Nga – Росія, Zelensky – Зеленський, Putin – путін, LHQ – ООН, quân đội – армія, nạn nhân – жертва, người dân – цивільне населення, người tị nạn – біженець*.

Далі було проведено повторну вибірку ключових слів через KeyBERT для кожного конкретного тексту (5 ключових слів на текст), а також перевірку на наявність заданих подій та персонажів у кожному тексті. На основі цих результатів за допомогою GPT-4 було виділено тематичні фрейми для кожної статті, як-от «Україна – жертва, яку підтримує світ», «Росія – агресор», «Є надія на розв'язання конфлікту», «Війна впливає на світову економіку» тощо. Ці фрейми було згруповано в тематичні групи, як-от: «Агресія та захист», «Гуманітарна криза», «Дипломатія та переговори», «Економічні наслідки».

Усі автоматично отримані результати ми також перевіряли вручну для верифікації, результат був задовільним. Більшість похибок припадала на визначення ключових слів, що зумовлено технічною недосконалістю. Групування й систематизація на абстрактному рівні за допомогою автоматизованих засобів та ручної верифікації збіглися повністю. Таким чином, можемо засвідчити ефективність індуктивного підходу з використанням доступних ресурсів.

5.3. Від кластерів до наративів

У другій частині дослідження було проведено кластеризацію усіх текстів за змістом. За допомогою K-means алгоритму тексти було розбито на кластери відповідно до семантичної подібності їхніх PhoBERT-ембедінгів. Обмеженням K-means є те, що від дослідника одразу вимагається визначитися з кількістю кластерів. Таким чином, занадто велика кількість може призвести до розпорошення результатів, а занадто мала – до можливої втрати важливих тематичних груп. Алгоритму було задано розбити корпус на 5 кластерів. За допомогою GPT проведено інтерпретацію кожного кластера, а саме – найменування кластера, визначення провідної теми, повторюваних ключових лексем та ймовірного ключового наративу. У результаті виділено такі кластери: «Військові дії», «Гуманітарна криза та евакуація», «Дипломатичні зусилля та міжнародна реакція», «Економічні наслідки та енергетична криза», «Інциденти та межі конфлікту». Наприклад, кластер «Військові дії» об'єднав тексти, які описують активні військові операції, обстріли населених пунктів та окупацію територій. Тема кластера: «Агресія та захист». Ключові слова: *tấn công – атака, chiến sự – воєнні дії, quân sự – військовий, tên lửa – ракета*. Провідний наратив: «Росія наступає – Україна захищається». Найбільш неоднозначним виявився кластер «Інциденти на межі конфлікту», який містив тексти про атаки на об'єкти цивільної інфраструктури, тероризм, порушення міжнародного права, а також повідомлення, що не стосувалися війни безпосередньо.

На практиці K-means виявився цінним інструментом для цього дослідження, однак метод вимагає попередньої вказівки кількості кластерів, що є викликом, коли заздалегідь невідомо, скільки різних типів наративів присутні в даних. З метою усунути ці прогалини було проведено додаткову кластеризацію за допомогою алгоритму HDBSCAN з параметром мінімального розміру кластера (min_cluster_size), який було варійовано від 10 до 3 залежно від щільності даних у векторному просторі. Алгоритм HDBSCAN виявився

оптимальним для цього завдання, оскільки він дозволяє автоматично визначати кількість кластерів на основі щільності точок (текстів) та природним чином ідентифікувати «шумові» спостереження (позначені як кластер -1), які не належать до жодної стійкої групи. Перший запуск алгоритму з параметром min_cluster_size=10 класифікував усі новини як шум (-1), що вказувало на недостатню щільність кластерів при такому гіперпараметрі. Після зниження порогу до min_cluster_size=3, алгоритм виявив 6 основних кластерів (позначених від 0 до 5) плюс категорія шуму (-1). Ці «шумові» тексти виявилися корисними для наративного аналізу, оскільки часто містили або дуже специфічні сюжети, або змішані фрейми, які не вписувалися в жоден із домінантних кластерів. Таким чином, HDBSCAN виконував роль інструмента для виявлення як ядрових, так і периферійних наративів, доповнюючи більш жорсткий K-means у змішаному кластеризаційному підході.

У результаті інтерпретації моделлю GPT кластерам присвоєно такі назви: «Глобальні економіко-гуманітарні наслідки війни», «Біженці та гуманітарна допомога», «Геополітичні наслідки війни», «Дипломатичне посередництво», «Нейтральні гравці у контексті війни», «Розвиток подій на фронті». До прикладу, кластер «Глобальні економіко-гуманітарні наслідки війни» об'єднав тексти, що розглядають не військові дії напряму, а вторинні наслідки війни, російсько-українська війна згадується як контекст, що впливає на ринки, торгівлю, виробництво та постачання продовольства, зокрема на в'єтнамський експорт, інфляцію, ціни на енергоносії. Провідні наративи: «Війна впливає на життя людей в усьому світі», «Війна призводить до негативних наслідків у світовій економіці».

До кластеру «шуму» в тому числі потрапили тексти, що не стосувалися війни, зокрема частина новин січня 2022 року. Тому було проведено додаткову класифікацію усіх текстів на ті, що пов'язані з війною, та непов'язані. Валідацію кластеризації здійснено вручну, результати задовільні.

Результати кластеризації за допомогою K-means та HDBSCAN показали суттєвий збіг у виявлених змістовних осях конфлікту, але з різними рівнями деталізації. Зокрема, в обох класифікаціях видно перевагу у в'єтнамських новинах висвітлення не безпосереднього аналізу військових дій, а впливу війни на місцеву та глобальну економіку, політику та гуманітарну сферу. Таким чином, застосування такої перехресної кластеризації слугує додатковою верифікацією отриманих результатів.

6. Висновки

Проведене пілотне дослідження продемонструвало ефективність обраної методології для автоматизованого виявлення в'єтнамських медіанаративів про російсько-українську війну за умов обмежених ресурсів. Обробка експериментального корпусу текстів за допомогою Underthesea, KeyBERT, PhoBERT / SimCSE, K-means, HDBSCAN та GPT дозволила виявити основні наративні вісі: військові дії, гуманітарна криза, дипломатія та глобальні наслідки. Ключовою знахідкою стало домінування гуманітарного фрейму у в'єтнамських медіа. В'єтнамські офіційні ЗМІ демонструють нейтральну позицію, акцентуючи на людських стражданнях та практичних наслідках для Азії та світу, що узгоджується з державною політикою «бамбукової дипломатії» – гнучкого балансування між великими державами.

Застосування абдуктивного підходу показало валідність використаних методів, адже в обох напрямах дослідження отримано зіставні результати. Крім того, ручна перевірка також підтвердила отримані дані та інтерпретації. Методологія також виявилася стійкою до ресурсно обмежених умов, зокрема в середовищі Google Colab, та цілком придатною для розширення корпусу на локальних GPU, Colab Pro або з використанням API від LLM. Отже, цей алгоритм може бути випробуваний для автоматичного виявлення наративів у великих обсягах даних.

Подальші дослідження можуть стосуватися зіставлення корпусів офіційних ЗМІ, соціальних мереж та експертних повідомлень, з визначенням NER для персонажів (PhoBERT+ViSOBERT) та динамічним трекінгом наративів через LSTM-моделі, із застосуванням цифрових гуманітарних наук для аналізу пропагандистських дискурсів. Цікавою може бути міжмовна компаративістика (в'єтнамськомовні vs. англомовні, україномовні медіа), однак для інших мов інструменти мають бути адаптовані, адже в цьому дослідженні завданням було розробити дієздатний алгоритм саме для в'єтнамської мови.

Дослідження підтверджує, що цифрові гуманітарні науки роблять наративний аналіз доступним для великих масивів даних. Методологія відкриває шлях до масштабних досліджень в'єтнамських дискурсів, де автоматичні інструменти та ШІ слугують прискорювачами та підсилювачами людського аналізу.

Подяка

Це дослідження й експеримент було проведено під наставництвом Наталії Чейлитко в рамках курсу «Великі мовні моделі для лінгвістів», організованого Єнським університетом імені Фрідріха Шиллера 11.02.-24.06.2025.

Primary Sources

*Báo tin tứ*c. https://baotintuc.vn

Chat GPT. https://chatgpt.com/

Google Colaboratory. *Colab FAQ: Resource limits*. https://research.google.com/colaboratory/faq.html

ParseHub. https://www.parsehub.com/

Underthesea – Vietnamese NLP Toolkit (2025). https://github.com/undertheseanlp/underthesea/tree/main

VoVanPhuc. (2024). *sup-SimCSE-Vietnamese-phobert-base.* https://huggingface.co/VoVanPhuc/sup-SimCSE-VietNamese-phobert-base

References

Abiodun, M. I., Absalom, E. E., Laith, A., Belal, A. & Jia, H. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Barthes, R. (1975). An introduction to the structural analysis of narrative. *New Literary History, 6(2*), 237–272. https://doi.org/10.2307/468419

Burch, R. (2024). Charles Sanders Peirce, *The Stanford Encyclopedia of Philosophy* (Summer 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.). https://plato.stanford.edu/archives/sum2024/entries/peirce/

Das, R., Chandra, A., Lee, I-Ta & Pacheco, M. L. (2024). Media framing through the lens of event-centric narratives. In *Proceedings of the 6th Workshop on Narrative Understanding*, 85–98. https://doi.org/10.18653/v1/2024.wnu-1.15

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43(4)*, 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Fairclough, N. (1989). *Language and power*. London: Longman.

Fairclough, N. (1995). *Critical discourse analysis: The critical study of language*. London: Longman.

German, F., Keith, B. & North, C. (2025). Narrative trails: A method for coherent storyline extraction via maximum capacity path optimization. In *Proceedings of the Text2Story'25 Workshop, Lucca (Italy), 10.04.2025*. https://doi.org/10.48550/arXiv.2503.15681

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harvard University Press.

Greimas, A. J. (1983). *Structural semantics: An attempt at a method*. University of Nebraska Press.

Grootendorst, M. (2020). *KeyBERT: Minimal keyword extraction with BERT. Python Package*. https://github.com/MaartenGr/KeyBERT

Hoang, T. H. & Dien Nguyen, A. L. (2022). The Russia-Ukraine war: Unpacking online pro-Russia narratives in Vietnam. *ISEAS Perspective*, *44*, 1–14. https://www.iseas.edu.sg/wp-content/uploads/2022/03/ISEAS_Perspective_2022_44.pdf

Hoang, T. H. (2025). Vietnam's mediascape amid the war in Ukraine: Between method and mayhem. In *Fulcrum*. https://fulcrum.sg/vietnams-mediascape-amid-the-war-in-ukraine-between-method-and-mayhem/

Keith, B. & Mitra, T. (2020). Narrative maps: An algorithmic approach to represent and extract information narratives. In *Proceedings of Text2Story 2020*. https://doi.org/10.48550/arXiv.2009.04508

Levi, E., Mor, G., Shenhav, S.R., & Sheafer, T. (2020). CompRes: A dataset for narrative structure in news. In *Proceedings of LREC 2020*. ArXiv, abs/2007.04874.

Mai, T. D. T. (2025). News framing on social media: A case study of Russia-Ukraine war narration on Facebook in Vietnam. *Keio Communication Review, 47(3)*, 33–61. https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/download.php/AA00266091-20250300-0033.pdf?file_id=187546

McInnes, L., & Healy, J. (2017). Accelerated hierarchical density clustering. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, 33–42. https://doi.org/10.1109/ICDMW.2017.12

Nguyen, D. Q., & Nguyen, A. T. (2020). PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1037–1042. https://doi.org/10.18653/v1/2020.findings-emnlp.92

Nguyen, N. T., Nguyen, D. V., Phan, K. T. K, & Nguyen, N. L.T. (2023). Abusive span detection for Vietnamese narrative texts. In *Proceedings of the 12th International Symposium on Information and Communication Technology (SOICT '23),* 471–478. https://doi.org/10.1145/3628797.3628921

Nguyen, S. T., Nguyen, N. L., Trang, T., Nguyen, T. H., Duong, T. T. P. & Tuan, N. (2021). Stock article title sentiment-based classification using PhoBERT. In *Proceedings of the 2nd International Conference on Human-centered Artificial Intelligence (Computing4Human 2021)*. https://ceur-ws.org/Vol-3026/paper25.pdf

Piper, A., So, R. J. & Bamman, D. (2021). Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 298–311. 10.18653/v1/2021.emnlp-main.26

Propp, V. (1968). *Morphology of the folktale* (2nd ed., L. Scott, Trans.). University of Texas Press.

To, M.S. (2022). Explaining the Vietnamese public's mixed responses to the Russia-Ukraine crisis. In *The Diplomat*. https://thediplomat.com/2022/03/explaining-the-vietnamese-publics-mixed-responses-to-the-russia-ukraine-crisis/

van Dijk, T. A. (1998). *Ideology: A multidisciplinary approach*. London: Sage.

van Dijk, T. A. (2008). *Discourse and context: A sociocognitive approach*. Cambridge, New York: Cambridge University Press.

Резюме

Мусійчук Вікторія

ПІЛОТНИЙ ЕКСПЕРИМЕНТ З АВТОМАТИЧНОГО ВИЯВЛЕННЯ В'ЄТНАМСЬКИХ МЕДІАНАРАТИВІВ ПРО РОСІЙСЬКО-УКРАЇНСЬКУ ВІЙНУ ПРИ ОБМЕЖЕНИХ РЕСУРСАХ

Постановка проблеми. Наративи у медіатекстах відіграють ключову роль у формуванні громадської думки щодо міжнародних конфліктів, зокрема російсько-української війни. Системне вивчення природи формування в'єтнамських медіанаративів є актуальним як з лінгвістичного погляду, так і для застосування результатів для вироблення ефективних стратегій міжнародної комунікації. Традиційний наративний аналіз неефективний для великих корпусів через ресурсомісткість, особливо для низькоресурсних мов як в'єтнамська (відсутність анотованих датасетів, складна токенізація морфем, обмежений доступ до багаторівневих даних). Попередні дослідження обмежені якісним дискурс-аналізом соцмереж, без автоматизованих методів NLP для масштабування.

Мета статті. Розробити та протестувати гібридну методологію автоматизованого видобування наративів з в'єтнамських медіатекстів за умов обмежених ресурсів, поєднуючи класичну наратологію з цифровими гуманітарними методами (NLP, кластеризація), для ідентифікації подієцентричних осей (події, персонажі, фрейми) та їхньої інтерпретації.

Методи дослідження. Пілотний експеримент на корпусі 160 новин з Báo tin tức, зібраних ParseHub, токенізований Underthesea. Абдуктивний підхід (Burch, 2024): 1) індуктивний – KeyBERT+PhoBERT/SimCSE-Vietnamese (ембедінг, ключові фрази), GPT-4 (групування на події/персонажів/теми); 2) дедуктивний – K-means/ HDBSCAN+PhoBERT/SimCSE-Vietnamese (ембедінг, кластеризація), GPT-4 (інтерпретація). Ручна верифікація результатів.

Основні результати дослідження. У першому напрямі експерименту за допомогою KeyBERT виявлено ключові фрази, які далі було розподілено за наративними мітками (події, персонажі), згруповано у тематичні групи та наративні фрейми за допомогою GTP. У другому напрямі проведено паралельну кластеризацію двома інструментами: K-means та HDBSCAN. Виявлені кластери інтерпретовано та виділено наративні фрейми й ключові терміни. У результаті зіставлення двовекторного підходу виявлено збіг у видобутих змістовних осях та наративних фреймах. Зокрема, в обох напрямах очевидна перевага висвітлення у в'єтнамських новинах не безпосереднього аналізу військових дій, а впливу війни на місцеву та глобальну економіку, політику та гуманітарну сферу.

Висновки і перспективи. Дослідження продемонструвало ефективність обраної методології для автоматизованого виявлення в'єтнамських медіанаративів про російсько-українську війну за умов обмежених ресурсів. Застосування абдуктивного підходу показало валідність методів, адже в обох напрямах дослідження отримано корелюючі результати. Методологія виявилася стійкою до ресурсно обмежених умов, зокрема в середовищі Google

Colab, та придатною для розширення корпусу. Подальші дослідження можуть стосуватися зіставлення різних корпусів, застосування NER для персонажів та динамічного трекінгу наративів через LSTM-моделі, сприяючи аналізу пропагандистських дискурсів.

Ключові слова: в'єтнамська мова, медіа, наративи, PhoBERT, кластеризація, російсько-українська війна.

Abstract

Musiichuk Viktoriia

PILOT EXPERIMENT ON AUTOMATIC DETECTION OF VIETNAMESE MEDIA NARRATIVES ABOUT THE RUSSIA-UKRAINE WAR UNDER LIMITED RESOURCES

Background. Narratives in media texts play a crucial role in shaping public opinion on international conflicts, including the Russia–Ukraine war. Systematic investigation of how Vietnamese media narratives are constructed is relevant both from a linguistic perspective and for developing effective strategies of international communication. Traditional narrative analysis is inefficient for large corpora due to its resource intensity, especially for low-resource languages such as Vietnamese (lack of annotated datasets, complex morpheme tokenization, limited access to multi-layered data). Existing studies are largely confined to qualitative discourse analysis of social media and do not employ scalable NLP-based automation.

Purpose. The aim of the article is to develop and test a hybrid methodology for the automated extraction of narratives from Vietnamese media texts under limited computational resources, combining classical narratology with digital humanities methods (NLP, clustering) in order to identify event-centric narrative axes (events, characters, frames) and provide their interpretation.

Methods. The study presents a pilot experiment on a corpus of 160 news items from Báo tin tức, collected via ParseHub and tokenized with Underthesea. An abductive approach (Burch, 2024) was implemented along two complementary strands: (1) an inductive strand using KeyBERT + PhoBERT / SimCSE-Vietnamese (text embeddings, keyphrase extraction) and GPT-4 (grouping into events, characters, and themes); (2) a deductive strand using K-means / HDBSCAN + PhoBERT / SimCSE-Vietnamese (embeddings, clustering) and GPT-4 (cluster interpretation). All automatic outputs were subjected to manual verification.

Results. In the first strand, KeyBERT was used to extract keyphrases, which were subsequently mapped onto narrative labels (events, characters), aggregated into thematic groups and narrative frames with the assistance of GPT-4. In the second strand, parallel clustering was performed with K-means and HDBSCAN. The resulting clusters were interpreted and associated with narrative frames and core lexical items. Comparison of the two-vector approach revealed convergence in the extracted semantic axes and narrative frames. In both strands, Vietnamese news were found to prioritise coverage of the war's impact on local and global economies, politics, and the humanitarian sphere over detailed analysis of military operations.

Discussion. The study demonstrates the effectiveness of the proposed methodology for automated detection of Vietnamese media narratives about the Russia–Ukraine war under limited resources. The abductive design proved methodologically valid,

as both strands produced consistent and mutually reinforcing results. The workflow is robust in resource-limited environments such as Google Colab and is scalable to larger corpora. Future research may include systematic comparison of different corpora, integration of NER for character extraction, and dynamic narrative tracking using LSTM-based models, thereby contributing to the analysis of propagandistic discourses.

Keywords: Vietnamese language, media, narratives, PhoBERT, clustering, Russia–Ukraine war.

Відомості про автора

Мусійчук Вікторія Анатоліївна, кандидат філологічних наук, старший науковий співробітник, завідувач відділу Азіатсько-Тихоокеанського регіону Інституту сходознавства ім. А. Ю. Кримського НАН України, e-mail: v.musiychuk@nas.gov.ua

Musiichuk Viktoriia, Ph.D. Philology, Senior researcher, A. Krymskyi Institute of Oriental Studies National Academy of Sciences of Ukraine, Head of Asia-Pacific Department, e-mail: v.musiychuk@nas.gov.ua

ORCID 0000-0002-4456-4487

Надійшла до редакції 30 листопада 2025 року
Прийнято до друку 25 грудня 2025 року