УДК 004.8:007.5:159.9:303.77 **doi: 10.26906/SUNZ.2024.1.141**

В. В. Нарожний, В. С. Харченко

Національний аерокосмічний університет “Харківський авіаційний інститут”, Харків, Україна

МЕТОД СЕМАНТИЧНОГО АНАЛІЗУ ДАНИХ ДЛЯ ВИЗНАЧЕННЯ МАРКЕРНИХ СЛІВ ПРИ ОБРОБЛЕННІ РЕЗУЛЬТАТІВ ОЦІНКИ ВІЗИТОРІВ В ІНТЕРАКТИВНОМУ МИСТЕЦТВІ

Анотація. Предметом дослідження є поглиблений семантичний аналіз даних, що базується на інтеграції методологій латентного розподілу Діріхле (LDA) та двонаправленого кодувального представлення з трансформаторів (BERT). Це дослідження зосереджується на обробленні текстових даних, зокрема, оцінок відвідувачів інтерактивного мистецтва, для визначення слів-маркерів, які виділяють ключові емоційні та тематичні елементи. Мета: поглибити розуміння досвіду та сприйняття відвідувачами інтерактивних мистецьких інсталяцій шляхом визначення значущих слів-маркерів за допомогою комбінованого підходу LDA та BERT. Це комплексування має на меті охопити як загальний тематичний зміст, так і нюансований контекст зворотного зв'язку. Завдання: збір та попередня обробка текстових даних – оцінок відвідувачів, що складається з етапів токенізації, нормалізації та лематизації з впровадження LDA для виокремлення поширених тем із зібраних даних, що надає уявлення про основні теми, присутні у відгуках відвідувачів; інтеграція BERT для аналізу контекстуальних нюансів і виведення глибших значень з окремих слів у відгуках; поєднання результатів LDA та BERT для створення комплексного розуміння текстових даних, фокусуючись на виявленні найбільш значущих слів-маркерів. Досягнуто такі результати: виконано успішне виокремлення ключових тем з оцінок відвідувачів за допомогою LDA, що дозволило виявити широкі тематичні категорії, присутні у відгуках; запропоновано підхід глибокого навчання BERT, що забезпечив нюансовані контекстні вбудовування, підкреслюючи конкретні емоції та настрої, висловлені відвідувачами; здійснено інтеграцію результатів LDA та BERT, що надало багатий набір слів-маркерів, які ефективно відображають суть досвіду та сприйняття відвідувачами інтерактивного мистецтва; покращено точність і глибина аналізу у визначенні ключових емоційних і тематичних елементів, про що свідчить узгодженість і релевантність слів-маркерів відносно оцінок відвідувачів. Висновки: інтеграція LDA та BERT для семантичного аналізу даних в інтерактивних мистецьких контекстах демонструє потужний підхід для розуміння складних відгуків відвідувачів. Цей метод забезпечує дворівневий аналіз, де LDA пропонує розуміння загальних тем, а BERT сприяє детальному контекстуальному розумінню. Дослідження успішно визначає конкретні слова-маркери, які ефективно передають суть вражень та оцінок відвідувачів. Ця методологія може бути корисною для художників, кураторів та дослідників у вимірюванні публічної рецепції та покращенні інтерактивного мистецького досвіду. Адаптивність методології створює реальні перспективи її застосування в інших сферах, де потрібен детальний семантичний аналіз текстових відгуків.

Ключові слова: семантичний аналіз даних, обробка природної мови, прихований розподіл Діріхле, двонаправлені кодерні представлення з трансформаторів, інтерактивне мистецтво, аналіз емоційної реакції.

Вступ. Підхід до аналізу вербальних відгуків

Останніми роками розвиток цифрових платформ і соціальних мереж полегшив збір величезної кількості текстових відгуків від відвідувачів. Однак суб'єктивний і часто складний характер цих відгуків створює виклик: як ми можемо отримати значущі висновки з таких різноманітних і нюансованих даних? Саме тут відбувається перетин семантичного аналізу даних та обробки природної мови (NLP).

Семантичний аналіз даних передбачає вивчення тексту для виявлення закономірностей, тем і настроїв, які можуть бути неочевидними з першого погляду. Він дозволяє витягти глибші значення і зв'язки в тексті, забезпечуючи більш тонке розуміння змісту. У цьому дослідженні для вирішення цього завдання використано два інструменти NLP - латентний розподіл Діріхле (LDA) та двонаправлене кодування за допомогою трансформаторів (BERT).

LDA, популярна техніка тематичного моделювання, використовується для виявлення спільних тем і напрямків у зібраних відгуках. Він допомагає класифікувати текстові дані за окремими темами, надаючи структурований огляд відповідей відвідувачів. Однак LDA має свої обмеження в розумінні контексту та відтінків значень слів і фраз. Щоб вирішити цю проблему, ми інтегрували BERT - сучасну модель представлення мови, відому своєю здатністю вловлювати контекст слів у тексті, враховуючи слова, що стоять до і після речення. Підхід BERT, заснований на глибокому навчанні, дозволяє проводити більш контекстуальний аналіз відгуків, визначаючи "слова-маркери" - ключові терміни, які мають важливе значення або сенс, пов'язані з досвідом відвідувачів. Поєднання LDA та BERT забезпечує комплексний інструментарій для аналізу текстових відгуків від інтерактивних мистецьких інсталяцій, сценарії яких проаналізовано в [1, 2]. Результатом є глибше розуміння досвіду відвідувачів, їхніх емоційних реакцій та елементів інсталяцій, які резонують з ними найсильніше.

Дана стаття пропонує новий підхід до аналізу відгуків відвідувачів в інтерактивних мистецьких інсталяціях, використовуючи сильні сторони як LDA, так і BERT. Таким чином, маємо на меті надати художникам, кураторам і дослідникам цінну інформацію про вплив і сприйняття інтерактивних творів мистецтва, тим самим сприяючи розвитку дискурсу в галузі цифрового інтерактивного мистецтва.

1. State of the Art

Інтеграція семантичного аналізу даних в інтерактивне мистецтво зазнала значного прогресу за останні роки. Нові дослідження демонструють поєднання традиційного мистецтвознавчого аналізу з сучасними обчислювальними методами, зокрема, у

розумінні взаємодії відвідувачів та зворотного зв'язку в мистецьких інсталяціях.

Досягнення в галузі соціально-семантичного мережевого аналізу [3] заглиблюються в дуальність соціальних і семантичних мереж, підкреслюючи, як ці мережі впливають одна на одну. Їхнє розуміння соціально-семантичного мережевого аналізу підкреслює потенціал поєднання соціальних взаємодій із семантичним аналізом, що є актуальним для інтерактивного мистецтва, де зворотний зв'язок з відвідувачами охоплює як соціальний, так і індивідуальний досвід. Венскович і Норт [4] досліджують семантичну взаємодію у високорозмірних даних за допомогою підходу інтерактивного семантичного дослідження. Їхня методологія, зосереджена на кластеризації та проекції у двовимірному просторі, тісно пов'язана з процесом переробки складних відгуків відвідувачів у зрозумілі тематичні структури. Чжоу та ін. [5] пропонують ієрархічну модель взаємодії між модальностями для візуально-текстового аналізу настроїв, підкреслюючи важливість семантичної та емоційної взаємодії між різними модальностями. Це дослідження збігається з метою нашого дослідження - проаналізувати відгуки відвідувачів, які часто поєднують візуальні та текстові елементи. Штуббеманн, Дюрршнабель і Реффлінгхаус [6] обговорюють семантичний аналіз погляду у віртуальній і доповненій реальності, підкреслюючи важливість розуміння візуального сприйняття в інтерактивних умовах. Їхні висновки сприяють нашому розумінню того, як відвідувачі взаємодіють з інтерактивним мистецтвом, забезпечуючи основу для аналізу візуальної уваги та взаємодії. Сучасні дослідження в галузі семантичного аналізу даних, зокрема в контексті інтерактивного мистецтва, демонструють тенденцію до інтеграції різноманітних обчислювальних методів для розуміння складних людських взаємодій та досвіду. Методології варіюються від мережевого аналізу та інтерактивного дослідження даних до передового аналізу настроїв, кожна з яких дає цінну інформацію про відгуки відвідувачів та їхню залученість. Ці досягнення не лише збагачують наше розуміння досвіду відвідувачів в інтерактивному мистецтві, але й відкривають нові шляхи для міждисциплінарних досліджень у мистецтві, соціальних науках та комп'ютерному аналізі.

В даній статті розробляється та досліджується метод семантичного аналізу даних для визначення маркерних слів при обробленні результатів оцінки візиторів в інтерактивному мистецтві і забезпечення більш точного вибору. Ця задача розв’язуються в такій послідовності:

- на першому етапі відбувається попередня обробка даних. На цьому етапі з даних відфільтровуються зайві дані;

- на другому етапі відбувається обробка даних алгоритму LDA. На цьому етапі дані групуються на кластери;

- на третьому етапі відбувається пост обробка результату роботи алгоритму LDA алгоритмом BERT. На цьому етапі кластери з даними проходять додаткову обробку для виявлення більш точних маркерних слів для кожного кластера.

2. Попередня обробка даних

Попередня обробка даних є критично важливим етапом методології, який гарантує, що зібрані дані є чистими, структурованими та готовими до аналізу. Цей етап включає кілька кроків для перетворення необроблених текстових відгуків у формат, придатний для семантичного аналізу за допомогою латентного розподілу Діріхле (LDA) та двонаправленого кодувального представлення з трансформаторів (BERT). Ось детальний опис етапів попередньої обробки даних:

2.1. Нормалізація. Нормалізація в обробці тексту передбачає перетворення текстових даних у послідовний формат для полегшення точного аналізу [7]. У контексті підготовки даних для семантичного аналізу за допомогою C# на платформі .NET нормалізація зазвичай включає нормалізацію регістру, видалення діакритичних знаків (акцентів) і стандартизацію варіацій у тексті. Нормалізація регістру.

Мета. Забезпечити однорідність тексту, оскільки під час аналізу великі та малі літери одного і того ж слова повинні розглядатися як ідентичні.

Реалізація на C#. Використання методу ToLower() для перетворення всього тексту в нижній регістр.

2.2. Видалення діакритичних знаків (наголосів). *Мета*. У мовах з наголошеними символами часто буває корисно стандартизувати ці символи до їх базової форми, особливо коли семантичний аналіз не розрізняє наголошені та ненаголошені символи [8].

Реалізація на C#: Використання простору імен System.Text.Normalization та методу string.Normalize() для декомпозиції символів з діакритичними знаками, а потім видалення непідкреслених символів.

2.3. Стандартизація варіацій тексту. *Мета.* Обробка варіацій у тексті, які слід обробляти однаково, наприклад, синоніми, регіональні відмінності у написанні (наприклад, американська та британська англійська) або специфічна термінологія, що використовується як взаємозамінна [9].

Реалізація на C#. Створення словника варіацій та їхніх стандартизованих форм, потім заміна в тексті.

2.4. Токенізація. Токенізація – це важливий процес в аналізі тексту, під час якого текст розбивається на менші одиниці, як правило, слова або фрази. Ці одиниці, відомі як токени, є основою для подальшого аналізу, такого як синтаксичний, структурний або семантичний аналіз [10]. У C# токенізація може бути виконана за допомогою різних методів, залежно від складності тексту та вимог аналізу.

2.5. Розширена токенізація. *Мета.* Обробка складних текстових структур, таких як речення з розділовими знаками, скороченнями або іншими мовами з іншими правилами токенізації.

2.6. Видалення стоп-слів. Видалення загальних слів, які не роблять істотного внеску в зміст тексту, таких як "та", "і" і "або". Цей крок зменшує обсяг даних, що підлягають обробці, і допомагає зосередитися на словах, які мають більшу семантичну вагу [11].

2.7. Ідентифікація стоп-слів. Користувацький список стоп-слів залежно від конкретних потреб аналізу або тематики тексту може знадобитися створити власний список стоп-слів. Цей список може

включати специфічні для даної області терміни, які часто зустрічаються, але не є інформативними. Використання HashSet для зберігання стоп-слів є більш ефективним, ніж список або масив, оскільки забезпечує швидший час пошуку.

2.8. Лематизація. *Мета.* лематизація зводить слова до їхньої лематизованої форми, що передбачає складніший лінгвістичний аналіз для правильного приведення слова до його словникової форми. При цьому враховується частина мови, час та інші граматичні фактори [11]. *Реалізація на C#.* Ефективна реалізація лематизації вимагає всеосяжної лінгвістичної бази даних і складних алгоритмів, доступ до яких зазвичай можна отримати через бібліотеки НЛП. Однією з таких бібліотек є Stanford NLP, яку можна інтегрувати з додатками на C#.

2.9. Перетворення даних. Перетворення даних в обробці тексту - це важливий етап, на якому попередньо оброблений текст перетворюється в числовий формат, придатний для аналізу, особливо в моделях машинного навчання [11]. Цей процес включає в себе кілька ключових методів, кожен з яких призначений для обробки різних аспектів текстових даних.

2.10. Векторизація. *Мета.* Перетворення текстових даних у числові вектори, оскільки більшість алгоритмів потребують числових даних. *Метод.* Частота терміна, обернена до частоти документа (TF-IDF) – Подібний до BoW, але враховує частоту слова в усьому наборі даних, надаючи меншу вагу більш поширеним словам.

Завдяки ретельному виконанню цих етапів попередньої обробки дані ефективно перетворюються в чистий, структурований формат. Таке вдосконалення є важливим для подальших аналітичних етапів, щоб отримати точні та змістовні висновки з відгуків відвідувачів про інтерактивні мистецькі інсталяції.

3. Інтеграція LDA в обробку даних

Прихований розподіл Діріхле (Latent Dirichlet Allocation, LDA) - це популярна техніка моделювання тем, яка використовується в обробці природної мови (NLP) для виявлення абстрактних тем у колекції документів [12]. Впровадження LDA передбачає кілька ключових кроків, кожен з яких має важливе значення для вилучення значущих інсайтів з текстових даних. LDA ґрунтується на припущенні, що кожен документ є сумішшю різних тем і що кожна тема характеризується розподілом слів. Метою LDA є зворотне проектування цієї структури: враховуючи слова в документах, LDA намагається визначити набір тем, які, найімовірніше, згенерували б цю колекцію документів. Оцінка якості тем, згенерованих за допомогою LDA, є суб'єктивною, але дуже важливою. Для цього був вибраний наступний метод: Оцінка когерентності: Вимірює ступінь семантичної схожості між словами з високими показниками в темі. Вищі показники зв'язності зазвичай відповідають темам, які легше інтерпретувати.

Впровадження бібліотеки на основі Python, такої як Gensim для латентного розподілу Діріхле (LDA), у проекті на C# передбачає використання інтероперабельності між C# та Python. Цього можна досягти за допомогою таких інструментів, як Python.NET або IronPython. Python.NET є більш підходящим вибором для цього сценарію, оскільки він дозволяє C# взаємодіяти з Python та його бібліотеками напряму.

4. Інтеграція BERT в обробку даних

Двонаправлені кодові представлення з трансформаторів (BERT) – це метод, розроблений компанією Google для попереднього навчання NLP [13]. Інтеграція BERT в задачі обробки тексту включає кілька кроків, від початкового вибору моделі до кінцевої інтерпретації результатів. BERT призначений для розуміння контексту слова в реченні, дивлячись на слова, що стоять до і після нього. BERT вимагає вхідних даних у певному форматі.

Багато завдань можуть вирішуватися за допомогою попередньо навчених BERT-моделей безпосередньо або з додатковим налаштуванням. Такі бібліотеки, як Hugging Face's Transformers, забезпечують простий спосіб завантаження та використання цих моделей. Результати BERT можуть бути складними. Для задач вилучення ознак BERT надає вбудовування для кожного токена. Для задач класифікації вихідні дані з токена [CLS] можна подавати на додаткові шари, щоб отримати остаточну класифікацію.

5. Приклад практичного використання

Для ілюстрації зберемо набір даних, де кожен запис –опис емоцій відвідувачів картинної галереї.

Набір даних:

1. "Сьогодні я відчуваю себе надзвичайно щасливим!"
2. "Усередині мене глибокий смуток".
3. "Я киплю від гніву".
4. "Хвиля спокою накрила мене".
5. "Я здивований поворотом подій".

Кожне речення попередньо обробляється:

- приведення тексту до нижнього регістру;
- видалення розділових знаків та спецсимволів;
- розбиття тексту на слова;
- видалення стоп-слів (таких як "я", "є", "з");
- застосування лематизації.

Попередньо оброблені дані:

1. ["відчуваю", "абсолютно", "радісно", "сьогодні"];
2. ["глибоко", "смуток", "всередині"];
3. ["кипіння", "гнів", "несправедливість"];
4. ["хвиля", "спокій", "накритий"];
5. ["здивований", "поворот", "події"].

Реалізація LDA. Використовуючи LDA, ми визначаємо теми в цих описах емоцій. Ми встановили кількість тем 3. LDA може класифікувати дані наступним чином: Тема 1: ["радісний", "щасливий", "схвильований"] (позитивні емоції); Тема 2: ["сум", "горе", "печаль"] (негативні емоції); Тема 3: ["здивований", "шокований", "вражений"] ("емоції здивування").

Інтеграція BERT. Використовуючи BERT, ми генеруємо вставки для кожного попередньо обробленого речення, щоб вловити контекстні нюанси. Далі ми можемо використовувати ці вставки для таких завдань, як класифікація емоцій.

Встроювання BERT:

1. [0.85, -0.12, ...] (позитивні емоції);

2. [-0.76, 0.33, ...] (негативна емоція);
3. [-0,60, 0,29, ...] (негативна емоція);
4. [0,47, -0,15, ...] (позитивна емоція);
5. [0.22, 0.67, ...] (емоція здивування);

Поєднання результатів LDA та BERT. Після генерації тем LDA та вбудовувань BERT об'єднуємо ці результати, щоб отримати більш повне розуміння емоційного контексту кожного текстового запису. Для кожного речення ми пов'язуємо його з найбільш релевантною темою LDA і доповнюємо її контекстним розумінням, яке надають вбудовування BERT.

Комбінований аналіз:

1. Тема LDA: Позитивні емоції, BERT: [0,85, -0,12, ...] → Радісний, оптимістичний;
2. Тема ЛДА: Негативні емоції, BERT: [-0,76, 0,33, ...] → Сумний, меланхолійний;
3. Тема ЛДА: Негативні емоції, BERT: [-0,60, 0,29, ...] → Злий, розчарований;
4. Тема ЛДА: Позитивні емоції, BERT: [0.47, -0.15, ...] → Спокійний, миролюбний;
5. Тема ЛДА: Емоції здивування, BERT: [0.22, 0.67, ...] → Здивований, Заінтригований.

Об'єднані результати дають змогу глибше зрозуміти кожне речення. Теми LDA забезпечують широку категоризацію емоцій, тоді як вбудовування BERT пропонують нюансоване контекстно-залежне розуміння. Інтегруючи LDA і BERT, ми можемо ефективно аналізувати текстові дані, щоб виокремити як широкі тематичні елементи, так і тонкі контекстні нюанси емоцій людей. Такий підхід забезпечує більш глибоке і детальне розуміння, ніж будь-який з методів окремо, демонструючи силу поєднання різних технік NLP в аналізі текстів.

6. Обговорення

Основна методологія дослідження ґрунтується на синергетичному використанні LDA і BERT, двох сучасних технік NLP, кожна з яких робить свій унікальний внесок в аналіз текстового зворотного зв'язку. Модель LDA ефективно виокремлює широкі тематичні структури з відгуків, класифікуючи загальні настрої та теми, що переважають серед відвідувачів. На противагу цьому, роль BERT була ключовою в аналізі складних контекстуальних значень конкретних фраз і слів, що дозволило виявити слова-маркери, які інкапсулюють нюанси емоційних реакцій відвідувачів. Комбінований підхід забезпечив багатовимірне розуміння відгуків відвідувачів. Слова-маркери, визначені за допомогою цієї методології відобразили спектр реакцій від радості та здивування до роздумів і критики. Ці результати не лише підтверджують ефективність комплексного підходу, але й підкреслюють складність досвіду відвідувачів в умовах інтерактивного мистецтва. Для художників і кураторів ці висновки є досить важливими. Вони пропонують засновану на даних основу для розуміння залучення та реакції аудиторії. Це розуміння може вплинути на майбутні мистецькі творіння, дизайн виставок і навіть на кураторство інтерактивного досвіду, забезпечуючи глибший резонанс з аудиторією.

Поза межами інтерактивного мистецтва ця методологія має ширше застосування. Подібні підходи можна застосовувати в інших сферах, де розуміння суспільних настроїв і сприйняття має вирішальне значення, наприклад, в аналізі відгуків про продукт, аналізі настроїв у соціальних мережах та у тематичних дослідженнях у рамках якісних досліджень.

Дослідження визнає певні виклики. Обчислювальна інтенсивність BERT і необхідність ретельного налаштування параметрів у LDA є нетривіальними міркуваннями. Крім того, сфера дослідження була обмежена специфічним контекстом інтерактивного мистецтва, що може вплинути на узагальненість результатів. Майбутні дослідження можуть вивчити застосування цієї методології до різних наборів даних і контекстів для подальшого підтвердження її ефективності. Забігаючи наперед, можна сказати, що дослідження відкриває шляхи для включення більш складних методів NLP і вивчення семантичного аналізу, керованого ШІ. Удосконалення можуть включати аналіз зворотного зв'язку в реальному часі, крос-культурні порівняльні дослідження в інтерактивному мистецтві та інтеграцію мультимодального аналізу даних для включення візуального та слухового зворотного зв'язку поряд з текстовими даними.

Висновки

Досліджуючи сучасні методи аналізу тексту, ми заглибилися в синергетичну інтеграцію двох потужних технік NLP: прихованого розподілу Діріхле (LDA) та двонаправленого кодування за допомогою трансформаторів (BERT). Ця комбінація є значним кроком у галузі обробки природної мови, пропонуючи комплексний підхід до розуміння як широких тематичних структур, так і складних контекстуальних нюансів у текстових даних.

LDA довів свою ефективність у виявленні прихованої тематичної структури у великих текстових масивах, надаючи високорівневе уявлення про домінуючі теми. Його здатність розбивати величезні обсяги тексту на зрозумілі теми є безцінною для початкового розвідувального аналізу.

BERT робить наступний крок, аналізуючи текст на детальному рівні. Підхід, заснований на глибокому навчанні, до генерації вкладених слів фіксує тонкі контекстні значення слів на основі навколишнього тексту, що призводить до більш тонкого розуміння мови. Інтеграція LDA та BERT забезпечує можливість проведення дворівневого аналізу. У той час як LDA класифікує текст за ширшими темами, BERT забезпечує глибину, вловлюючи нюанси та складнощі мови, які можуть бути пропущені LDA. Цей комбінований підхід особливо ефективний у таких додатках, як аналіз настроїв, рекомендація контенту та емоційний аналіз.

Незважаючи на свою потужність, цей комплексний підхід не позбавлений проблем. Обчислювальна інтенсивність BERT, ретельне налаштування, необхідне для LDA, і потреба в надійних стратегіях попередньої обробки та валідації підкреслюють складність розширеного аналізу тексту.

Подальші дослідження і розробки можуть бути присвячені пошуку і аналізу метрик, які оцінюють якість визначення маркерних слів за допомогою

запропонованого та інших методів, а також розвиток і практичне впровадження напрацьованої методології для оброблення суджень-відповідей експертів, які надаються у вербальній формі, при розв'язанні важко формалізованих задач, зокрема, при оцінюванні безпеки [14].

СПИСОК ЛІТЕРАТУРИ

1. O. Golembovska, V. Kharchenko, I. Shostak, M. Danova, and O. Feoktystova. Assessing the Perception of Abstract Paintings with Elements of Augmented Reality, 11th IEEE DESSERT, Ukraine, 2020. DOI: 10.1109/DESSERT50317.2020.9125014.
2. O. Golembovska, V. Kharchenko, I. Shostak, M. Danova, O. Feoktystova, and V. Plietnov, Augmented Reality for the Abstract Paintings: Application Scenarios, Semantic Similarity Analysis and Case Study, *2019 10th IEEE Int. Conf. on IDAACS.: Technology and Applications*, Metz, France, 2019, pp. 1007-1011, DOI: 10.1109/IDAACS.2019.8924411.
3. N. Basov, R. Breiger, I. Hellsten. Socio-semantic and other dualities. Poetics. 2020. p.101433, DOI: 10.1016/j.poetic.2020.101433.
4. Pollux: Interactive Cluster-First Projections of High-Dimensional Data [Текст] / John E. Wenskovitch, C. North // 2019 IEEE Visualization in Data Science (VDS) - 2019. – pp.38-47, DOI: 10.1109/VDS48975.2019.8973381.
5. Visual-Textual Sentiment Analysis Enhanced by Hierarchical Cross-Modality Interaction [Текст] / Tao Zhou, Jiuxin Cao, Xueling Zhu, Bo Liu, Shancang Li // IEEE Systems Journal - 2021. – pp.4303-4314, DOI: 10.1109/jsyst.2020.3026879.
6. Neural Networks for Semantic Gaze Analysis in XR Settings / Lena Stubbemann, Dominik Dürrschnabel, R. Refflinghaus // ACM Symposium on Eye Tracking Research and Applications - 2021, DOI: 10.1145/3448017.3457380.
7. Multilingual Sequence Labeling Approach to solve Lexical Normalization / Divesh R. Kubal, Apurva Nagvenkar // 2021 The 7th Workshop on Noisy User-generated Text (W-NUT) - 2021. – p.457-464, DOI: 10.18653/v1/2021.wnut-1.51.
8. Proposed Natural Language Processing Preprocessing Procedures for Enhancing Arabic Text Summarization / Reda Elbarougy, G. M. Behery, Akram el Khatib // 2019. – p.39-57, DOI: 10.1007/978-3-030-34614-0_3.
9. The influence of preprocessing on text classification using a bag-of-words representation / Yaakov Hacohen-Kerner, Daniel Miller, Yair Yigal // PLoS ONE - 2020. – DOI: 10.1371/journal.pone.0232525.
10. Italian Text Categorization with Lemmatization and Support Vector Machines / F. Camastra, Gennaro Razi // 2020. – p.47-54, DOI: 10.1007/978-981-13-8950-4_5.
11. From Words to Numbers: Getting Started with Text Analysis for Applied Social Scientists [Текст] / Hyun Woo Kim, Hyejung Chang // BCRP (Business Communication Research and Practice) - 2020. – p.122-129, DOI: 10.22682/BCRP.2020.3.2.122.
12. A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura / S. Zhou, P. Kan, Qunying Huang, J. Silbernagel // Journal of Information Science - 2021. DOI: 10.1177/01655515211007724.
13. Neural Topic Models for Short Text Using Pretrained Word Embeddings and Its Application To Real Data / R. Murakami, B. Chakraborty. 2021 IEEE 4th Int Conf on Knowledge Innovation and Invention (ICKII) - 2021. – p.146-150, DOI: 10.1109/ICKII51822.2021.9574752.
14. Babeshko, I.; Leontiiev, K.; Kharchenko, V.; Kovalenko, A.; Brezhniev, E. Application of Assumption Modes and Effects Analysis to XMECA. In Theory and Engineering of Dependable Computer Systems and Networks; DepCoS-RELCOMEX 2021. Advances in Intelligent Systems and Computing; Springer: Cham, Switzerland, 2021; Volume 1389. DOI: 10.1007/978-3-030-76773-0_1.

Received (Надійшла) 21.12.2023
Accepted for publication (Прийнята до друку) 07.02.2024

Method of semantic data analysis for determining marker words in the processing of visitors' evaluation results in interactive art

V. Narozhnyi, V. Kharchenko

Abstract. The subject of the study is in-depth semantic data analysis based on the integration of the methodologies of latent Dirichlet distribution (LDA) and bidirectional encoding representation from transformers (BERT). This research focuses on processing textual data, in particular, visitors' evaluations of interactive art, to identify marker words that highlight key emotional and thematic elements. The goal is to deepen the understanding of visitors' experiences and perceptions of interactive art installations by identifying significant marker words using a combined LDA and BERT approach. This combination aims to capture both general thematic content and the nuanced context of feedback. Objectives: collection and preprocessing of textual data - visitor ratings, consisting of tokenization, normalization and lemmatization steps with the implementation of LDA to extract common themes from the collected data, providing insights into the main themes present in visitor feedback; integration of BERT to analyze contextual nuances and extract deeper meanings from individual words in the feedback; combining the results of LDA and BERT to create a comprehensive understanding of the textual data, focusing on identifying the most significant marker words. The following results were achieved successful extraction of key themes from visitors' ratings using LDA, which allowed us to identify broad thematic categories present in the reviews; a deep learning approach BERT was proposed, which provided nuanced contextual embeddings, emphasizing specific emotions and sentiments expressed by visitors; the results of LDA and BERT were integrated, which provided a rich set of marker words that effectively reflect the essence of the experience and perception of visitors to interactive art; the accuracy and depth of analysis in identifying key emotional and thematic elements was improved, as evidenced by the consistency and relevance of marker words in relation to visitors' ratings. Conclusions: The integration of LDA and BERT for semantic data analysis in interactive art contexts demonstrates a powerful approach for understanding complex visitor feedback. This method provides a two-level analysis, where LDA offers insights into general themes and BERT contributes to detailed contextual understanding. The study successfully identifies specific marker words that effectively capture the essence of visitors' impressions and ratings. This methodology can be useful for artists, curators, and researchers in measuring public reception and improving interactive art experiences. The adaptability of the methodology creates real prospects for its application in other areas that require a detailed semantic analysis of textual feedback.

Keywords: semantic data analysis, natural language processing, latent Dirichlet distribution, bidirectional coded representations from transformers, interactive art, emotional response analysis.