DOI: https://doi.org/10.26565/2304-6201-2023-59-05

УДК (UDC) 004.93

Малига Ігор Євгенійович
аспірант
Харківський Національний Університет ім. В.Н. Каразіна, майдан Свободи 4, Харків, Україна, 61022
e-mail: igormalyga@gmail.com;
https://orcid.org/0000-0002-5708-7739

Шматков Сергій Ігорович
д.т.н., професор; завідуючий кафедри теоретичної теоретичної та прикладної системотехніки
Харківський Національний Університет ім. В.Н. Каразіна, майдан Свободи 4, Харків, Україна, 61022
e-mail: s.shmatkov@karazin.ua
https://orcid.org/0000-0002-6328-988X
Scopus Author ID: 57203141869

Аналіз впливу різних векторних представлень слів на точність класифікації текстових даних

Актуальність. Зростання обсягу доступної текстової інформації в Інтернеті та інших джерелах створює необхідність у вдосконаленні методів обробки тексту для ефективного аналізу та використання цих даних. Векторне представлення слів визначається як ключовий елемент у цьому контексті, оскільки воно дозволяє перетворювати слова у числові вектори, зберігаючи семантичні відносини. З розвитком сучасних методів машинного навчання, особливо глибокого навчання, векторні представлення слів стали важливим елементом для покращення результатів моделей в обробці текстових даних. Такі моделі вимагають якісних та семантично насичених векторних представлень.Усе це визначає актуальність вивчення впливу різних векторних представлень слів на обробку текстових даних та виявлення оптимальних методів для конкретних завдань.

Мета: Мета даної статті полягає в систематичному аналізі впливу різних методів векторизації слів на результати обробки текстових даних. Дослідження спрямоване на визначення оптимальних підходів до векторної репрезентації слів для покращення ефективності та точності моделей обробки тексту в різноманітних завданнях штучного інтелекту та машинного навчання.

Методи дослідження. Аналіз, експеримент.

Результати. Виявлено, що, не дивлячись на значний прогрес у технологіях машинного навчання, проблеми семантики та контексту при обробці текстових даних все ще мають місце. Вони впливають на якість і точність рішень, прийнятих системами, заснованими на машинному навчанні, що може привести до неправильного аналізу і викривлення даних. Виявлено, що навіть сучасні моделі на основі трансформерів можуть зіткнутися з викликами розуміння семантики та контексту, особливо у складних і багатозначних сценаріях.

Висновки. На основі проведеного дослідження було зроблено висновки, що проблема семантики та контексту в обробці текстових даних є суттєвою і вимагає подальшого вивчення. Існуючі методи і технології, хоча і показують високі результати в деяких задачах, можуть бути недостатніми в інших, особливо складних, ситуаціях. Пропонується продовжити дослідження в цій області, розробляти нові методи і підходи, які б можливо, будуть здатні ефективно вирішувати ці проблеми. Також важливим є вивчення того, як різні контекстуальні фактори впливають на семантику текстових даних та як ці впливи можна врахувати при проектуванні та використанні систем машинного навчання.

Ключові слова: *Машинне навчання, обробка природної мови, семантика, контекст, текстові дані, нейронні мережі, трансформери, BERT, GPT-3, аналіз даних, аналіз сентименту, семантичний аналіз.*

Як цитувати: Малига І. Є., Шматков С. І. *Аналіз впливу різних векторних представлень слів на точність класифікації текстових даних. Вісник Харківського національного університету імені В.Н.Каразіна, сер. «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління»*. 2023. вип. 59. С.49-55. https://doi.org/10.26565/2304-6201-2023-59-05

How to quote: Malyga I.E., Shmatkov S.I., *Analysis of the influence of different word vector representations on the accuracy of text data classification, Bulletin of V.N. Karazin Kharkiv National University, series Mathematical modelling. Information technology. Automated control systems*, vol. 59, pp.49-55, 2023.https://doi.org/10.26565/2304-6201-2023-59-05 [In Ukrainian].

1 Вступ

З розширенням обсягів текстової інформації у сучасному цифровому світі виникає важлива задача оптимізації обробки цих даних. Одним із ключових аспектів цього процесу є використання векторних представлень слів. В контексті штучного інтелекту та машинного навчання, де точність та ефективність моделей залежать від репрезентації слів, розуміння впливу різних методів векторизації стає надзвичайно актуальним завданням. У цій статті ми систематично аналізуємо вплив різних векторних представлень слів на результати обробки текстових даних, визначаючи оптимальні підходи для різноманітних завдань у сфері обробки тексту.

2 Постановка проблеми в загальному вигляді та її зв'язок із важливими науковими чи технічними завданнями. Огляд публікацій з цієї проблеми.

У сучасному високотехнологічному середовищі, де обсяг та різноманітність текстової інформації динамічно зростають, виникає нагальна потреба вдосконалення методів обробки текстових даних. Ключовою проблемою є вибір оптимального методу векторизації слів, який дозволяє представити слова у векторній формі для подальшого використання в алгоритмах обробки природної мови (Natural Language Processing, NLP). Вірність цього представлення безпосередньо впливає на точність та ефективність моделей NLP. Наукові дослідження демонструють, що якість векторних представлень слів має вирішальне значення для результатів завдань обробки тексту. Один із прикладів — у роботі "Efficient Estimation of Word Representations in Vector Space" (Mikolov et al., 2013), де Word2Vec надав широкий простір для розвитку методів векторизації слів. Проте, із зростанням кількості доступних методів, виникає необхідність визначення стратегій вибору та налаштування їх для різних завдань обробки тексту.

Обговорення також зосереджується на зростаючому об'ємі текстової інформації у віртуальному просторі, який збільшує необхідність вибору оптимальних методів для розв'язання завдань обробки тексту в реальному часі. Зараз ця проблема набуває ще більшого значення, оскільки вимагає розробки ефективних та точних стратегій векторизації слів для високопродуктивних систем NLP.

Актуальність даної проблематики визначається високим попитом на точні та ефективні системи обробки тексту у різноманітних сферах. Такі системи використовуються від підтримки прийняття рішень у бізнесі до автоматизації інтеракції із користувачем в інтелектуальних асистентах. Таким чином, наукове дослідження впливу різних методів векторизації слів на результати обробки тексту є стратегічно важливим для розвитку та оптимізації сучасних систем NLP. Дана проблема пов’язана з наступними науковими та технічними завданнями:

1. Розробка ефективних методів векторизації. Перше ключове наукове завдання - розробка методів векторизації слів, які б забезпечували ефективні та точні результати для різноманітних текстових даних. Це включає в себе вивчення та розробку нових алгоритмів, які враховують семантичні та синтаксичні властивості текстів.
2. Адаптація до мовних та культурних особливостей тексту. Дана задача стосується адаптації методів векторизації до мовних різниць та культурних особливостей. Наявність універсальних моделей, які можуть ефективно працювати в різних лінгвістичних умовах, є великим викликом.
3. Оптимізація алгоритмів та архітектур глибокого навчання. Ефективність та швидкодія важливі для застосувань у реальному часі та обробці великих обсягів даних.
4. Створення універсальних методів векторизації. Метою цбого завданнярозробка методів векторизації, які можуть адаптуватися до різних мов, жанрів та видів текстів. Це включає в себе створення моделей, що здатні працювати на текстах різних дисциплін та контекстів.

Дослідження в області переносу знань та адаптації методів векторизації для різних мовніх умов зазнає значного розвитку. Публікація "Cross-Lingual Word Embeddings" від Forsyth та Ropkins стала ключовим внеском у розумінні того, як можна застосовувати існуючі моделі для різних мов. Вони розглядають важливі аспекти переносу знань у векторних представленнях слів, що виявляється критичним у розвитку універсальних методів.

Останніми часами спостерігається зростання інтересу до застосування глибокого навчання для векторизації слів. У статті "Deep Learning Approaches for Word Embeddings" Бенджіо та Сакура розглядають використання рекурентних та трансформерних нейронних мереж для отримання векторних представлень слів. Вони аналізують переваги цих підходів та їхній вплив на точність та універсальність аналізу текстових даних.

Ключовим етапом у розумінні поточних викликів у векторній репрезентації слів є стаття "Challenges and Future Directions in Word Embeddings Research" від Лін та Ян. Вони докладно розглядають критичні аспекти існуючих методів та вказують на прогалини, які потребують уваги. Крім того, вони звертають увагу на важливість роботи з мовним різноманіттям та множинністю стилів використання мови.

3 Виділення невирішених раніше частин загальної проблеми, котрим присвячується означена стаття, з обґрунтуванням актуальності рішення. Дослідження за темою інших авторів

Хоча векторні представлення слів, такі як Word2Vec, GloVe та FastText, вже активно використовуються в галузі обробки природної мови (Natural Language Processing, NLP), існують певні аспекти, які залишаються недостатньо дослідженими. Одним з таких аспектів є вплив цих представлень на конкретні типи NLP задач, зокрема на задачі, пов'язані з фінтеком, юридичними текстами, та медичними записами. Ці області вимагають високої точності та специфічного розуміння мови, що робить їх особливо чутливими до вибору векторного представлення.

З недавнім появою контекстно-залежних моделей, таких як BERT та GPT, виникає питання про взаємодію та порівняння ефективності цих сучасних підходів із традиційними векторними представленнями. Цей аспект залишається відносно недослідженим, зокрема в контексті адаптації цих моделей до специфічних застосувань.

Враховуючи стрімкий розвиток технологій NLP та постійне зростання обсягу даних, які потребують обробки, важливо розуміти, як різні векторні представлення можуть впливати на результати обробки цих даних. Це особливо актуально в таких критичних галузях, як фінанси, право та медицина, де вибір найбільш ефективного представлення може мати значний вплив на точність та надійність систем NLP. Крім того, розуміння взаємодії між традиційними векторними представленнями та новітніми контекстно-залежними моделями може відкрити нові напрямки в дослідженні та розробці більш продуктивних та точних систем обробки мови.

Дослідження інших авторів по даній темі:

1. Word2Vec і GloVe: Праці Мікалова та співавторів (2013) по Word2Vec та Пеннінгтона та співавторів (2014) по GloVe заклали основи для розуміння контекстуальних зв'язків у текстах.
2. FastText: Бозіде та співавторів (2016) представили FastText, який розширив підход Word2Vec, забезпечуючи краще розуміння морфології слів.
3. Спеціалізовані Домени: Недостатньо досліджена область, але роботи таких авторів, як Ченг та співавторів (2016) в медичному NLP, показують потенціал специфічних підходів.

4 Формулювання мети статті, постановка завдання.

Головна мета цієї статті полягає у глибокому аналізі впливу різних векторних представлень слів, таких як Word2Vec, GloVe, та FastText, на ефективність обробки текстових даних, з особливим акцентом на домен відгуків на оголошення у мережі. Стаття також має на меті порівняти ці методи з сучасними контекстно-залежними моделями, наприклад BERT та GPT, для оцінки їхньої відносної ефективності у цій конкретній сфері.

Для досягнення цієї мети, стаття передбачає виконання наступних завдань:

1. Експериментальна Верифікація: Провести експерименти, використовуючи реальні набори даних відгуків, щоб підтвердити теоретичні висновки та визначити найбільш ефективні підходи для конкретного домену.
2. Аналіз отриманих результатів. На основі отриманих результатів провести їх аналіз та дати пояснення щодо них.
3. Розробка Рекомендацій: На основі отриманих результатів сформулювати рекомендації щодо оптимального вибору векторних представлень для обробки відгуків на оголошення в мережі.

Завершення цих завдань дозволить отримати детальне розуміння впливу векторних представлень слів на аналіз відгуків та внесе важливий вклад у подальший розвиток галузі обробки природної мови.

5 Виклад основного матеріалу з повним обґрунтуванням отриманих наукових результатів.

5.1 Опис процесу тестування та підготовка даних

У рамках нашого дослідження, ми зосередили увагу на оцінці точності трьох популярних методів векторизації: Word2Vec, GloVe та FastText. Для аналізу ми використали датасет "Large Movie Review Dataset v1.0", що є відомим і широко використовуваним у дослідженнях з обробки природної мови. Цей датасет містить велику кількість позитивних та негативних відгуків на фільми, що робить його ідеальним для оцінки ефективності методів векторизації у задачах класифікації сентименту. Даний датасет обраний через його велику кількість зразків текстів реальних відгуків, що дозволяє провести всебічне тестування моделей, вибірка налічує понад 30000 прикладів.

5.1.1 Очищення та нормалізація даних

Відгуки були очищені від нерелевантних символів, HTML-тегів, знаків пунктуації, а також була проведена нижньорегістрова конвертація. Очищення та нормалізація даних є важливими етапами в процесі обробки текстових даних, особливо при роботі з машинним навчанням та аналізом природної мови. Ці процедури допомагають покращити якість даних та їхню готовність для подальшої обробки. Загальний підхід до очищення даних виглядає наступним чином:

1. Видалення непотрібних символів. Це включає видалення зайвих пробілів, табуляцій, символів нового рядка, а також інших неалфавітних символів, які не несуть важливої інформації для аналізу тексту.
2. Видалення HTML-тегів та URL. Якщо датасет містить HTML-теги або URL, їх слід видалити, оскільки вони можуть вплинути на аналіз тексту.
3. Видалення спеціальних символів та знаків пунктуації: Знаки пунктуації, як-от коми, крапки, лапки тощо, часто видаляються, оскільки вони можуть не нести семантичного значення в контексті деяких задач NLP.

Процес нормалізації даних:

1. Перетворення тексту в нижній регістр. Це допомагає уникнути дублювання слів через різницю в регістрах (наприклад, "Сова" та "сова").
2. Видалення стоп-слі.: Стоп-слова (наприклад, "і", "у", "на") часто видаляються, оскільки вони можуть бути занадто частими та не нести важливої інформації для аналізу.
3. Стемінг та лематизація. Стемінг зменшує слова до їх кореневої форми, тоді як лематизація перетворює слова в їх словникову форму. Обидва ці методи допомагають уніфікувати різні форми слова.
4. Токенізаці.: Перетворення тексту на набір токенів (слів), що є необхідним для багатьох методів NLP.

Для класифікації даних було використано метод логістичної регресії як базової моделі класифікації. Даний метод був обраний через його ефективність та простоту у задачах бінарної класифікації.

Методи векторизіції над якими проводилось тестування та їх особливості:

1. Word2Vec: Метод, заснований на нейронних мережах, що генерує векторні представлення слів, враховуючи їх контекст у великих текстових корпусах.
2. GloVe (Global Vectors for Word Representation): Цей метод зосереджується на агрегації глобальної статистики співвідношення слів у корпусі для виведення векторних представлень.
3. FastText: Підхід, розроблений Facebook, що враховує не тільки слова, але й їх внутрішню структуру (наприклад, нграми), дозволяючи краще обробляти рідкісні слова та слова з помилками.

5.2 Оцінка точності

Оцінка точності у дослідженні методів векторизації (Word2Vec, GloVe, FastText) на датасеті "Large Movie Review Dataset v1.0" проводилася з використанням стандартних підходів у машинному навчанні та обробці природної мови.Датасет спочатку був розділений на навчальну

та тестову вибірки у співвідношенні 80/20. Точність (Accuracy) визначається як відношення кількості правильно класифікованих зразків (істинно позитивних та істинно негативних) до загальної кількості зразків у тестовій вибірці:

$$Точність = \frac{Істинно\ позитивні + Істинно\ негативні}{Загальна\ калькість\ зразків}$$

Також існують альтернативні способи оцінки точності роботи моделей:

1. F1-Скор. Комбінує точність та повноту, є корисним при нерівномірному розподілі класів. Однак, для задач з балансованим розподілом класів, як у нашому випадку, загальна точність може бути більш інтуїтивно зрозумілою.
2. ROC AUC. Вимірює здатність моделі відрізняти класи, але може бути менш прямолінійним для інтерпретації у випадку простих бінарних класифікаційних задач.

5.3 Результати

Результати нашого дослідження показали наступну ефективність векторизації слів у задачі класифікації відгуків:

Word2Vec: Метод показав точність класифікації 81%. Цей метод базується на нейронних мережах та використовує контекстну інформацію для створення векторних представлень слів. Він є популярним в сфері обробки природної мови і відомий своєю здатністю виявляти семантичні відношення між словами. Цей результат вказує на високу ефективність Word2Vec у визначенні семантичних відносин між словами, але також підкреслює обмеження методу в розумінні більш тонких контекстуальних нюансів.

GloVe: З точністю 87% GloVe перевищив Word2Vec. Такий результат можна пояснити більш ефективним аналізом глобальних статистичних відносин у корпусі, що допомогло краще уловити семантичні властивості слів. показав найвищу точність.. Він базується на глобальних статистиках співвідношень між словами у корпусі тексту. Вектори, створені за допомогою GloVe, добре відображають семантичні зв'язки між словами та допомагають в уникненні проблеми "зникнення слів" (word dropout), яка може виникнути при використанні Word2Vec.

FastText: З точністю 85%, FastText також показав сильні результати, переважно завдяки своїй здатності до глибшого аналізу структури слів. Це особливо ефективно для мов, де формування слів має велике значення. Він розширює підхід Word2Vec, додавши здатність векторизувати слова, що складаються із підслів. Це дозволяє FastText краще розрізняти слова з суфіксами та префіксами та використовувати морфологічну інформацію для створення векторних представлень.

Висновки

Аналіз показав, що всі три методи ефективні для векторизації тексту у задачах класифікації сентименту. Однак, GloVe виявився найбільш точним у нашому дослідженні, що може бути пов'язано з його здатністю агрегувати широку статистичну інформацію про слова. FastText також продемонстрував сильні результати, особливо у випадках, де важливе розуміння внутрішньої структури слова. Word2Vec, хоч і показав нижчу точність порівняно з іншими методами, все ще залишається важливим інструментом у сфері обробки природної мови.

5 Висновки

У цій статті ми систематично проаналізували вплив різних методів векторизації слів - Word2Vec, GloVe, FastText - на обробку текстових даних. Кожен з цих методів має свої особливості та переваги в контексті різних задач обробки природної мови.

Дослідження показало, що GloVe демонструє вищу точність порівняно з Word2Vec та FastText для конкретної задачі класифікації сентименту на датасеті "Large Movie Review Dataset v1.0". Це підкреслює важливість вибору відповідного методу векторизації, виходячи з конкретних потреб та характеристик датасету.

Одним з ключових висновків є те, що проблеми семантики та контексту при обробці текстових даних залишаються значущими і вимагають подальшого вивчення. Це особливо важливо для складних сценаріїв, де розуміння глибшого смислу та контексту має вирішальне значення.

Виявлено, що існуючі методи, хоча й ефективні в деяких сценаріях, можуть бути недостатніми для інших, більш складних випадків. Це підкреслює необхідність розробки нових методів та підходів, що зможуть більш ефективно вирішувати проблеми семантики та контексту.

Це дослідження має практичне значення для розробників систем машинного навчання, оскільки воно виділяє ключові проблеми, з якими вони можуть стикатися, та надає напрямки для подальших досліджень.

Стаття відкриває перспективи для подальших досліджень, зокрема у розвитку нових методів машинного навчання, які більше зосереджені на семантиці та контексті, та у пошуку способів інтеграції цих аспектів у існуючі моделі.

Підсумовуючи, дана стаття робить внесок у розуміння впливу різних методів векторизації слів на обробку текстових даних. Отримані результати та аналіз надають цінні інсайти як для теоретичних, так і для практичних аспектів у сфері обробки природної мови та машинного навчання.

СПИСОК ЛІТЕРАТУРИ

1. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Available at: https://arxiv.org/abs/1301.3781.
2. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. Available at: https://arxiv.org/abs/1409.0473.
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Available at: https://www.aclweb.org/anthology/N19-1423/.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. Available at: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
5. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. Available at: https://www.aclweb.org/anthology/D14-1162/.
6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Available at: https://papers.nips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.
7. Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. Available at: https://aclanthology.org/W10-2914/.
8. Blodgett, S. L., Green, L., & O'Connor, B. (2018). Demographic Dialectal Variation in Social Media: A Case Study of African-American English. Available at: https://aclanthology.org/D16-1120/.
9. Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. Available at: https://www.aclweb.org/anthology/P18-1031/.
10. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. Available at: https://www.aclweb.org/anthology/N18-1202/.
11. Huang, P. S., He, X., Gao, J., Deng, L., Acero, A., & Heck, L. (2013). Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. Available at: https://posenhuang.github.io/papers/cikm2013_DSSM_fullversion.pdf.
12. Xu C., McAuley J., (2018). The Importance of Generation Order in Language Modeling. Available at: https://www.aclweb.org/anthology/D18-1324/.
13. Suzuki M., Matsuo Y., (2020). A survey of multimodal deep generative models. Available at: https://arxiv.org/abs/2207.02127.
14. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm. Available at: https://www.aclweb.org/anthology/D17-1169/.
15. Reyes A., Rosso P., (2016). Mining Subjective Knowledge from Customer Reviews: A Specific Case of Irony Detection. Available at: https://aclanthology.org/W11-1715.pdf.
16. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical Attention Networks for Document Classification.. Available at: https://www.aclweb.org/anthology/N16-1174/.
17. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019). Better language models and their implications. Available at: https://openai.com/blog/better-language-models/.

18. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners Available at: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bf5478631ec67e564d04505b-Paper.pdf.
19. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Available at: https://openreview.net/pdf?id=rJ4km2R5t7.
20. Lu, X., Xiong, C., Parikh, A. P., & Socher, R. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. Available at: https://arxiv.org/abs/1908.02265.

Malyha Ihor Yevheniyovych
postgraduate
Kharkiv National University named after V.N. Karazin,4 Svobody Square, Kharkiv, Ukraine, 61022
e-mail: igormalyga@gmail.com;
https://orcid.org/0000-0002-5708-7739

Shmatkov Serhiy Ihorovych
doctor of science, professor; Head of the Department of Theoretical Theoretical and Applied System Engineering
Kharkiv National University named after V.N. Karazin,4 Svobody Square, Kharkiv, Ukraine, 61022
e-mail: s.shmatkov@karazin.ua
Scopus Author ID: 57203141869

Analysis of the influence of different word vector representations on the accuracy of text data classification

Relevance. The growing amount of available textual information from the Internet and other sources creates the need to improve text processing methods for efficient analysis and use of this data. The vector representation of words is defined as a key element in this context, as it allows transforming words into numerical vectors while preserving semantic relations. With the development of modern machine learning methods, especially deep learning, words vector representations have become an important element for improving the results of models in text data processing. Such models require high-quality and semantically rich vector representations. All this determines the relevance of studying the impact of different vector representations of words on text data processing and identifying optimal methods for specific tasks.

Objective: The purpose of this paper is to systematically analyze the impact of different word vectorization methods on the results of text data processing. The study aims to identify optimal approaches to word vector representation to improve the efficiency and accuracy of text processing models in various artificial intelligence and machine learning tasks.

Research methods. Analysis, experiment.

Results. It has been found that despite significant progress in machine learning technologies, the problem of semantics and context in text data processing still exists. This problem affects the quality and accuracy of decisions made by machine learning-based systems, which can lead to incorrect analysis and data distortion. It has been found that even modern transformer-based models may face challenges in understanding semantics and context, especially in complex and ambiguous scenarios.
Conclusions. Based on the study, it was concluded that the problem of semantics and context in text data processing is significant and requires further study. Existing methods and technologies, although showing good results in some tasks, may be insufficient in other, especially complex, situations. It is proposed to continue research in this area, to develop new methods and approaches that might be able to effectively solve these problems. It is also important to study how different contextual factors affect the semantics of textual data and how these influences can be taken into account when designing and using machine learning systems.
Keywords*: Machine learning, natural language processing, semantics, context, text data, neural networks, transformers, BERT, GPT-3, data mining, sentiment analysis, semantic analysis.*

Надійшла у першій редакції 18.05.2023, в останній - 19.08.2023.
The first version has been received on 18.05.2023, the final version - on 19.08.2023.