УДК 004.048

М. Є. Завальнюк

ДОСЛІДЖЕННЯ МОДЕЛЕЙ ДЛЯ АНАЛІЗУ НАСТРОЮ ТЕКСТУ

Вінницький національний технічний університет

Анотація

Дане дослідження присвячено аналізу можливостей моделей для аналізу настрою тексту. Шляхом огляду та порівняльного аналізу традиційних методів, моделей машинного навчання та підходів глибокого навчання, визначено ключові аспекти цих моделей у виявленні та класифікації настрою тексту. Зокрема, дослідження виявляє переваги та недоліки кожного підходу, а також розглядає їхнє застосування в різних галузях, що допомагає зрозуміти їхню реальну вартість та ефективність.

Ключові слова: аналіз настрою тексту, машинне навчання, обробка природної мови, штучний інтелект, емоційний аналіз.

Abstract

This study analyzes the capabilities of models for text sentiment analysis. Through a review and comparative analysis of traditional methods, machine learning models, and deep learning approaches, the key aspects of these models in detecting and classifying text sentiment are identified. In particular, the study identifies the advantages and disadvantages of each approach, as well as examines their application in various industries, which helps to understand their real value and effectiveness.

Keywords: text sentiment analysis, machine learning, natural language processing, artificial intelligence, emotional analysis.

Вступ

Аналіз настрою, також відомий як sentiment analysis, є методом обробки природної мови, який визначає положення тексту (позитивне, негативне або нейтральне). Аналіз настрою дозволяє обробляти великі кількості даних у режимі реального часу. Наприклад, можна автоматизувати бізнес-процеси та отримати інсайти про прийняття рішень на основі даних, аналізуючи тисячі відгуків на продукти, звернень до служби підтримки або твітів [1].

Метою даного дослідження є проведення комплексного аналізу моделей для аналізу настрою тексту з метою визначення їхньої ефективності та придатності для практичного застосування. Основними завданнями роботи є: огляд існуючих методів аналізу настрою тексту, порівняльний аналіз традиційних методів, та моделей машинного навчання, а також визначення можливих сфер застосування цих моделей.

Традиційні методи

Традиційні методи аналізу емоцій використовують ручний або напівавтоматичний аналіз тексту, щоб визначити емоційний тон. Хоча ці методи не такі масштабовані, як сучасні методи машинного навчання, вони можуть бути корисними для дослідницьких цілей або для невеликих наборів даних.

Ось декілька поширених традиційних методів Sentiment Analysis [2]:

- Лексичний аналіз: цей метод використовує словники слів, які пов'язані з певними емоціями (наприклад, "щасливий", "сумний", "злий"). Текст аналізується на предмет наявності цих слів, і їхня кількість використовується для класифікації настрою тексту.
- Правила на основі знань: цей метод використовує набір правил, які визначають, які слова або фрази вказують на певний настрій. Текст аналізується на предмет наявності цих слів або фраз, і правила використовуються для класифікації настрою тексту.
- Гібридні методи: ці методи поєднують два або більше традиційних методів Sentiment Analysis для покращення точності.

Їхні переваги та недоліки наведені у таблиці 1

Таблиця 1 – Переваги і недоліки різних типів чат-ботів

Переваги	Недоліки
Ці методи зазвичай прості для розуміння та реалізації.	Ці методи можуть бути непрактичними для великих наборів даних.
Легко зрозуміти, як ці методи класифікують текст, оскільки вони ґрунтуються на чітких правилах або словниках.	Точність цих методів залежить від якості словників, які використовуються.
Ці методи можуть бути ефективними навіть з невеликими наборами даних.	Ці методи можуть потребувати значної ручної роботи для створення правил або словників.

Сучасні моделі машинного навчання

Сучасні моделі Sentiment Analysis, засновані на машинному навчанні, зазвичай мають значно більшу точність і масштабованість, ніж традиційні методи. Ці моделі навчаються на великих наборах даних тексту й етикеток настрою, щоб виявити закономірності, які пов'язують слова та фрази з певними емоціями.

BERT (Bidirectional Encoder Representations from Transformers): ця модель, розроблена компанією Google, використовує нейронну мережу Transformer для навчання на великих наборах даних тексту. BERT може генерувати векторні представлення слів, які враховують контекст, у якому вони використовуються. Ці векторні представлення потім можна використовувати для різних завдань обробки природної мови, включаючи Sentiment Analysis [3]. BERT використовує бідирекційний підхід, що означає, що вона враховує як лівий, так і правий контекст для кожного слова в тексті. Модель попередньо навчається на великій кількості непозначеного тексту, використовуючи завдання “замаскованого мовлення” (masked language modeling) та “наступного речення” (next sentence prediction). Після попереднього навчання BERT може бути доналаштована лише одним додатковим вихідним шаром для створення сучасних моделей для різних завдань, таких як відповіді на питання та лінгвістичне виведення. BERT показала нові результати на одинадцяти завданнях обробки природної мови, включаючи підвищення GLUE-показника до 80,5% (покращення на 7,7%), точності MultiNLI до 86,7% (покращення на 4,6%) та F1-показника відповідей на питання SQuAD v1.1 до 93,2 (покращення на 1,5 пункту) та SQuAD v2.0 до 83,1 (покращення на 5,1 пункту) [3]. BERT враховує контекст на рівні слів, що дозволяє їй краще розуміти семантику тексту.

GloVe (Global Vectors for Word Representation): ця модель, розроблена компанією Stanford University, навчається на великих наборах даних тексту, щоб генерувати векторні представлення слів. GloVe використовує статистичний метод, який називається "ко-тренуванням", щоб навчити слова, подібні за значенням, мати подібні векторні представлення. Ці векторні представлення потім можна використовувати для різних завдань обробки природної мови, включаючи Sentiment Analysis [4]. GloVe навчається на агрегованих глобальних статистиках співвідношень між словами в корпусі тексту. Він використовує статистику співвідношень між словами, щоб створити векторні представлення слів. Результати навчання GloVe відображають цікаві лінійні підструктури в просторі векторів слів.

Якщо порівнювати ці дві моделі то можна отримати наступне:

1. Розмір словникового словника:
   - BERT: зазвичай має значно більший словниковий запас порівняно з GloVe. Це через те, що BERT навчається на великому корпусі тексту, включаючи Інтернет, що дозволяє йому враховувати більше слів та виразів.
   - GloVe: як правило, має менший словниковий запас, оскільки він побудований на основі статистики з великого корпусу тексту, але не має можливості враховувати контекстуальні зв'язки.
2. Якість векторів:
   - BERT: вектори, отримані з BERT, зазвичай мають вищу якість через його здатність до розуміння контексту. Це дозволяє BERT уникнути проблем з полісемією та дисамбігуацією, що можуть виникнути з GloVe.
   - GloVe: хоча GloVe також надає якісні вектори для слів, вони можуть не бути такими точними у випадках, коли слова мають різні значення в різних контекстах.
3. Використання у завданнях Sentiment analysis:
   - BERT: широко використовується у багатьох завданнях аналізу мови, таких як кла-

сифікація тексту, розпізнавання іменованих сутностей, машинний переклад та багато інших завдань.

- GloVe: також використовується в аналізу мови, але зазвичай його вектори використовуються для завдань, які не вимагають глибокого розуміння контексту, таких як пошук схожих слів або кластеризація тексту.

У кінцевому підсумку вибір між BERT і GloVe залежить від конкретної задачі та вимог проекту. BERT зазвичай віддається перевага в завданнях, де важлива глибока розуміння контексту, тоді як GloVe може бути більш підходящим у завданнях, де потрібні прості векторні представлення слів.

Висновки

Дане дослідження дозволило провести комплексний аналіз моделей для аналізу настрою тексту, виявивши їхні сильні та слабкі сторони, а також потенційні сфери застосування.

Традиційні методи, такі як лексичний аналіз та правила на основі знань, прості у реалізації та інтерпретації, але мають обмеження щодо масштабованості та точності, особливо при роботі з великими обсягами даних.

Моделі машинного навчання, такі як BERT та GloVe, демонструють значно вищу точність та ефективність, особливо з урахуванням контексту та семантики тексту. Вони здатні обробляти великі набори даних та виявляти складні закономірності, що робить їх придатними для різноманітних завдань аналізу настрою.

Вибір оптимальної моделі залежить від конкретного завдання, розміру та типу даних, а також вимог до точності та інтерпретованості результатів.

Потенційні сфери застосування моделей аналізу настрою:

- Маркетинг та реклама: аналіз відгуків клієнтів, оцінка ефективності рекламних кампаній, моніторинг бренду.
- Соціальні медіа: аналіз громадської думки, виявлення трендів, управління репутацією.
- Фінанси: аналіз новин та прогнозування ринкових тенденцій, оцінка ризиків.
- Служба підтримки клієнтів: автоматизація обробки звернень, визначення рівня задоволеності клієнтів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Getting Started with Sentiment Analysis using Python. Hugging Face – The AI community building the future. URL: https://huggingface.co/blog/sentiment-analysis-python (дата звернення: 06.05.2024).
2. Shivanandhan M. What is Sentiment Analysis? A Complete Guide for Beginners. freeCodeCamp.org. URL: https://www.freecodecamp.org/news/what-is-sentiment-analysis-a-complete-guide-to-for-beginners/ (дата звернення: 06.05.2024).
3. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. URL: https://arxiv.org/abs/1810.04805 (дата звернення: 06.05.2024).
4. GloVe: Global Vectors for Word Representation. ACL Anthology. URL: https://aclanthology.org/D14-1162/ (дата звернення: 06.05.2024).

Завальнюк Максим Євгенович — студент групи 3акітр-23м, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: mezgoodle@gmail.com.

Zavalniuk Maksym Yev. — Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, email : mezgoodle@gmail.com.