УДК 004.89 : 004.912

Б. О. ЧЕРЕС
В. С. ОЗЕРАНСЬКИЙ

ОСОБЛИВОСТІ ЗАСТОСУВАННЯ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ АНАЛІЗУ ТА РОЗПІЗНАВАННЯ ТОНАЛЬНОСТІ ТЕКСТОВИХ КОМЕНТАРІВ

Вінницький національний технічний університет

Анотація

Запропоновано підхід до підвищення ефективності інформаційної технології аналізу та розпізнавання тональності текстових коментарів на основі комбінованого застосування методів векторних представлень слів, тонкого налаштування трансформерних моделей та каскадної маршрутизації за рівнем упевненості. Наведено особливості побудови методу автоматичного розширення тонального словника української мови за принципом семантичної подібності векторних представлень слів для охоплення сленгової лексики та неологізмів, відсутніх у базовому словнику, та визначено підходи до формування каскадної моделі класифікації на основі тонкого налаштування трансформерних моделей для точного розпізнавання тональності коментарів, які на лексичному рівень класифікує як неоднозначні. Обґрунтовано доцільність використання механізму каскадної маршрутизації на основі порогу впевненості лексичної оцінки для адаптивного підвищення обчислювальної ефективності технології аналізу тональності.

Ключові слова: аналіз тональності, обробка природної мови, інформаційні технології, семантична подібність, каскадна архітектура, BERT, трансформерні моделі, тональний словник.

Abstract

An approach is proposed to improve the effectiveness of information technology for analysing and recognising the tone of text comments, based on the combined application of methods involving vector representations of words, fine-tuning of transformer models, and cascaded routing based on confidence levels. The paper outlines the features of a method for automatically expanding the Ukrainian tone dictionary based on the principle of semantic similarity of word vector representations to cover slang and neologisms which are absent from the base dictionary, and approaches are defined for forming a cascaded classification model based on fine-tuning Transformer models for the accurate recognition of the tone of comments that are classified as ambiguous at the lexical level. The feasibility of using a cascade routing mechanism based on a confidence threshold for lexical assessment to adaptively improve the computational efficiency of tone analysis technology has been substantiated.

Keywords*: sentiment analysis, natural language processing, information technology, semantic similarity, cascade architecture, BERT, transformer models, sentiment lexicon.*

Вступ

Постійне зростання обсягів зоденної текстової комунікацій в мережі Інтернет та поширення платформ для обміну думками зумовлюють потребу в ефективних методах автоматизованого опрацювання тексту. Одним із таких напрямів є аналіз тональності (sentiment analysis) — це підгалузь обробки природної мови (Natural Language Processing або ж NLP), що полягає у автоматичному визначенні суб'єктивного та емоційного ставлення автора до певного об'єкта чи події. Аналіз тональності активно застосовується для аналізу відгуків, репутації брендів і суспільних настроїв. Водночас наявні підходи у цій сфері мають низку обмежень: лексикон-базовані методи недостатньо враховують контекст, класичні алгоритми машинного навчання потребують ретельного налаштування ознак, а методи глибокого навчання вимагають значних обчислювальних ресурсів, що особливо актуально для української мови

Метою цього дослідження є підвищення ефективності інформаційної технології аналізу та розпізнавання тональності текстових коментарів на основі комбінованого застосування методів векторних представлень слів, тонкого налаштування трансформерних моделей та каскадної маршрутизації за рівнем упевненості.

Результати дослідження

Аналіз існуючих програмних засобів для розпізнавання тональності текстів показав, що переважна більшість із них мають суттєві обмеження. Найчастіше вони спираються або на статичні тональні словники, або на узагальнені моделі, навчені на англомовній основі і не адаптовані до україномовного чи ситуативного контекстів. Наприклад, VADER (Valence Aware Dictionary and sEntiment Reasoner) використовує спеціально складений словник із 7 500 слів із фіксованими ваговими коефіцієнтами тональності. Однак, він не може бути автоматично оновленим при появі нової лексики та сленгу [1]. Для забезпечення прийнятної точності для низки поширених мов можуть бути використані Хмарні API , як наприклад Google Natural Language API та Amazon Comprehend. Проте варто зазначити, що їх основним недоліком є залежність від стороннього сервісу, що робить непридатним їх застосування у закритих інформаційних системах [2]. Окрім цього, жоден із розглянутих інструментів не пропонує механізму автоматичного поповнення словника та економного розподілу обчислювальних ресурсів між очевидними та неоднозначними випадками класифікації.

Попередня обробка текстових даних. Якість розпізнавання тональності безпосередньо залежить від ефективності попередньої обробки тексту. Стандартний процес обробки включає розбиття тексту на елементарні одиниці (токени), видалення зайвих слів, нормалізацію (приведення до нижнього регістру, усунення спеціальних символів), приведення слів до їхньої основної форми. У текстах поширених в соціальних мереж дуже важливим є робота з неформальною лексикою, скороченнями, абревіатурами, емодзі та сленгом, оскільки вони напряму впливають на емоційну тональність тексту.

За результатами досліджень, якість попередньої обробки тексту суттєво впливає на точність кінцевої моделі. Правильно виконана нормалізація тексту підвищує точність класифікації на 7–15% для класичних методів машинного навчання та на 3–8% для методів глибокого навчання [1, 2]. Попри це, виникають труднощі при обробці багатомовних текстів, що є характерним явищем для сучасного україномовного інтернет середовища. Для вирішення подібних завдань оптимальним є застосування моделей ідентифікації мов (language identification), що попередньо визначають мову кожного фрагменту тексту й обирають відповідний конвеєр обробки.

Класичні методи машинного навчання для аналізу тональності. Серед класичних методів машинного навчання найпоширеніші для виконання задач класифікації тональності варто виділити метод опорних векторів (SVM) та баєсовий наївний класифікатор. Метод SVM демонструє ефективну роботу в умовах високовимірного простору ознак, характерного для векторів TF-IDF. Це дозволяє йому будувати найкращу границю для розділення різних класів текстів, навіть при умові малої кількості навчальних даних. Для задач бінарної класифікації тональності (позитивна / негативна) SVM із ядром радіально-базисної функції досягає точності 82–88% на стандартних тестах. Баєсів наївний класифікатор, хоч і спрощений, демонструє конкурентоспроможні результати завдяки ефективному використанню умовних ймовірностей і є особливо придатним за малого обсягу навчальних даних [3, 4].

Однак, головна проблема цих методів полягає у тому, що вони не здатні враховувати контекстні залежності між словами. Вектор TF-IDF відображає частотні характеристики тексту, але не несе семантичної інформації про відношення між словами та їхній порядок у реченні. Це є критичним для розпізнавання заперечень та прихованого наміру, зазвичай при сарказмі. Цей недолік зумовив стрімкий розвиток підходів на основі глибокого навчання, здатних моделювати контекстуальні залежності.

Архітектури глибокого навчання. Двонаправлені мережі LSTM (Bi-LSTM) обробляють послідовність слів одночасно у прямому та зворотному напрямках, що дозволяє враховувати контекст сусідів порчу для кожного слова. Коли до цього додати механізм уваги, Bi-LSTM досягає точності 88–92% на англомовних датасетах порівняно з SVM (82–88%) [5, 6]. Але у порівняні з ними, найкращі результати демонструють трансформерні архітектури на базі BERT (Bidirectional Encoder Representations from Transformers). Вони використовують механізм MHAM (Multi-Head Attention Mechanism) для кодування довгострокових залежностей між словами незалежно від їхньої позиції в реченні і здатні досягти точності 93–95% на англомовних текстах [6, 7].

Цей підхід також працює добре з українськими текстами. Моделі XLM-RoBERTa та Ukr-RoBERTa можуть розпізнавати тон тексту з точністю понад 91%. Це краще, ніж більшість інших методів, тому ці моделі є хорошим вибором для складних випадків Проте навіть найточніша трансформерна модель не розв'язує проблему статичності словникових ресурсів, яка є першопричиною низького покриття

лексикон-орієнтованих методів.

Метод автоматичного розширення тонального словника на основі векторних представлень слів. Існуючі словники мають велике обмеження: вони статичні. Словники укладаються вручну і не охоплюють сленгові вирази, неологізми та розмовну лексику, які постійно виникають при комунікації в інтернеті. Для розв'язання цієї проблеми пропонується метод автоматичного розширення базового тонального словника української мови на основі семантичної подібності векторних представлень слів, що адаптує підхід до індукції тональних лексиконів через поширення оцінок (label propagation) у графі семантичної подібності [9] до україномовного контенту.

Базовий словник формується на основі існуючих лексикографічних розробок [10], доповнених ручною розміткою. Векторні представлення слів навчаються моделлю fastText [11] на корпусі україномовних текстів, що поєднує енциклопедичні джерела з текстами соціальних мереж; вибір fastText зумовлений урахуванням підслівної (субсловесної) інформації, що покращує якість векторних представлень та рідковживаних словоформ і дозволяє оцінювати навіть слова, яких не було в навчальному корпусі. Для кожного слова, відсутнього в базовому словнику, обчислюється косинусна подібність до k найближчих слів базового словника у векторному просторі. Тональна оцінка нового слова формується як зважена за подібністю сума оцінок цих k найближчих сусідів. Якщо максимальна подібність до жодного слова базового словника не перевищує заданого порогу, слово залишається неоціненим, що автоматично сигналізує лексиконному рівню про недостатню впевненість і слугує додатковою підставою для ескалації коментаря на трансформерний рівень.

Інтеграція спеціального словника до системи аналізу настрою тексту. Цей словник використовується на першому рівні системи аналізу, а саме спочатку ідентифікуються фрагменти тексту, написані різними мовами. Наступним кроком спеціальний словник із коригувальними коефіцієнтами для підсилювальних і пом'якшувальних слів та частки заперечення «не» обчислює загальний показник тональності коментаря. Якщо значення цього показника перевищує заданий поріг упевненості, результат класифікації формується одразу, без залучення моделі глибокого навчання. Лише ті коментарі, які отримали неоднозначну оцінку за словником, тобто показник близький до нуля, містять суперечливі сигнали або велику кількість слів, яких немає у словнику, — передаються на другий рівень аналізу: спеціальну модель XLM-RoBERTa або Ukr-RoBERTa. Для додаткового зниження обчислювальних витрат другого рівня доцільним є застосування дистильованої версії моделі за технікою knowledge distillation [12], навченої на м'яких метках базової трансформерної моделі, що дозволяє зберегти 95–97% точності за значно нижчої обчислювальної вартості.

Перевага запропонованого підходу при реалізації в інформаційній технології аналізу та розпізнавання тональності текстових коментарів над існуючими аналогами визначається синергетичним ефектом трьох компонентів. Векторні представлення слів усувають проблему статичності тонального словника, охоплюючи сленгову лексику та неологізми, відсутні в базовому вручну укладеному лексиконі. Трансформерна модель забезпечує точну контекстну класифікацію неоднозначних випадків на основі довгострокових семантичних залежностей між словами у реченні. Механізм каскадної маршрутизації замикає контур ефективності, спрямовуючи на ресурсомісткий трансформерний рівень лише ті коментарі, лексиконна оцінка яких є неоднозначною. Жоден із розглянутих існуючих засобів (VADER, TextBlob, хмарні API) не поєднує автоматичне розширення словникового покриття, контекстну точність трансформерних моделей та економний розподіл обчислювальних ресурсів одночасно в єдиній технології, адаптованій до специфіки україномовного контенту.

Висновки

Розроблення інформаційної технології аналізу та розпізнавання тональності текстових коментарів є актуальним напрямком досліджень, що має значний потенціал для підвищення точності й обчислювальної ефективності автоматизованого опрацювання україномовного контенту та наближення автоматизованих систем до рівня кваліфікованого людського модератора, здатного розпізнавати сленг, іронію та контекстуальні відтінки висловлювань. Застосування методів обробки природної мови дозволяє створювати адаптивні системи класифікації, які динамічно охоплюють нову лексику та коректно реагують на зміни в мовленнєвій практиці користувачів соціальних мереж.

Основними перевагами запропонованої інформаційної технології є автоматичне охоплення сленгової лексики та неологізмів на основі семантичної подібності векторних представлень слів,

завчасне виявлення неоднозначних випадків через контроль порогу впевненості лексиконної оцінки та економний розподіл обчислювальних ресурсів між лексиконним і трансформерним рівнями каскаду залежно від реальної складності вхідного коментаря. Проте, для успішної реалізації необхідно вирішити ряд викликів, пов'язаних із формуванням розміченого корпусу україномовних коментарів достатнього обсягу для тонкого налаштування трансформерної моделі, калібруванням порогів косинусної подібності та впевненості каскаду відповідно до специфіки різних доменів коментарів, а також підтвердженням запропонованого методу розширення словника на реальній вибірці коментарів, що містять код-світчинг і регіональні діалектні особливості.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. – 2nd ed. – Cambridge University Press, 2020. – 449 p. DOI: https://doi.org/10.1017/9781108639286
2. Minaee S., Kalchbrenner N., Cambria E., Nikzad N., Chenaghlu M., Gao J. Deep Learning–Based Text Classification: A Comprehensive Review // ACM Computing Surveys. – 2021. – Vol. 54, No. 3. – P. 1–40. DOI: https://doi.org/10.1145/3439726
3. Pang B., Lee L., Vaithyanathan S. Thumbs up? Sentiment Classification Using Machine Learning Techniques // Proceedings of EMNLP 2002. – 2002. – P. 79–86. DOI: https://doi.org/10.3115/1118693.1118704
4. Sebastiani F. Machine Learning in Automated Text Categorization // ACM Computing Surveys. – 2002. – Vol. 34, No. 1. – P. 1–47. DOI: https://doi.org/10.1145/505282.505283
5. Hochreiter S., Schmidhuber J. Long Short-Term Memory // Neural Computation. – 1997. – Vol. 9, No. 8. – P. 1735–1780. DOI: https://doi.org/10.1162/neco.1997.9.8.1735
6. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT 2019. – 2019. – P. 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423
7. Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., Stoyanov V. Unsupervised Cross-lingual Representation Learning at Scale // Proceedings of ACL 2020. – 2020. – P. 8440–8451. DOI: https://doi.org/10.18653/v1/2020.acl-main.747
8. Коник М. І. Порівняльний аналіз підходів та методів визначення тональності тексту в контексті опрацювання відгуків мешканців міста // Вісник Херсонського національного технічного університету. – 2025. – Т. 2, № 2(93). – С. 186–192. DOI: https://doi.org/10.35546/kntu2078-4481.2025.2.2.23
9. Оленич І. Я., Притула М. М., Сінькевич О. В., Хамар О. О. Система автоматичного визначення тональності тексту // Електроніка та інформаційні технології. – 2021. – Вип. 15. – С. 16–23. DOI: https://doi.org/10.30970/eli.15.2
10. Hamilton W. L., Clark K., Leskovec J., Jurafsky D. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora // Proceedings of EMNLP 2016. – 2016. – P. 595–605. DOI: https://doi.org/10.18653/v1/D16-1057
11. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching Word Vectors with Subword Information // Transactions of the Association for Computational Linguistics. – 2017. – Vol. 5. – P. 135–146. DOI: https://doi.org/10.1162/tacl_a_00051
12. Sanh V., Debut L., Chaumond J., Wolf T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter // arXiv preprint. – 2019. URL: https://doi.org/10.48550/arXiv.1910.01108

Черес Богдан Олегович – студент групи 2КН-25м, факультету інтелектуальних інформаційних технологій та автоматизації, кафедра комп`ютерних наук, Вінницький національний технічний університет, Вінниця, e-mail: bogcheres@gmail.com

Озеранський Володимир Сергійович – кандидат технічних наук, доцент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.

Bohdan O. Cheres – student of group 2KN-25m, Faculty of Intelligent Information Technologies and automation, Department for Computer Sciences, Vinnytsia National Technical University, Vinnytsia, e-mail: bogcheres@gmail.com

Volodymyr S. Ozeransky – Candidate of Technical Sciences (Ph.D.), Associate Professor at the Department of Computer Science, Vinnytsia National Technical University, Vinnytsia.