УДК 004.82
DOI: 10.15587/2313-8416.2018.135069

ДОСЛІДЖЕННЯ МЕТОДІВ АНАЛІЗУ ВІДГУКІВ ПРО ТОВАРИ МАГАЗИНІВ ЕЛЕКТРОНІКИ

© О. В. Вечур, О. М. Сподарець

Робота присвячена вивченню методів аналізу відгуків про товари магазинів електроніки. Предметом дослідження є відгуки про товари. Метою роботи є аналіз методів обробки природної мови в контексті задачі аналізу відгуків. Методом дослідження є комп'ютерне та математичне моделювання.
В роботі були розглянуті різні класи методів аналізу відгуків про товари магазинів електроніки, в якості практичної реалізації було проведено порівняння результатів передбачення. Результати дослідження мають застосування при аналізі відгуків будь-якого магазину
Ключові слова: *обробка природної мови, обчислювальні алгоритми, аналіз даних, комп'ютерна лінгвістика*

1. Вступ

Комп'ютерний аналіз тексту на природній мові активно розвивається в останні роки багатьма колективами. Доступні сьогодні обчислювальні потужності дозволяють застосовувати для обробки великих масивів документів широкий клас математичних методів, що сприяють ефективному вирішенню завдань пошуку, класифікації, кластерного аналізу, виявлення прихованих закономірностей в даних та ін.

З кожним роком загальний світовий обсяг електронної торгівлі збільшується на 10–20 %. За прогнозами експертів, до кінця 2018 року близько 18 % від усіх роздрібних продажів буде відбуватися в Інтернеті, а до 2040 року цей показник досягне 95 %. Збільшення кількості електронної торгівлі, особливо у сфері роздрібних продажів, нерозривно пов'язане зі збільшенням кількості користувачів Інтернет-магазинів.

Важливу роль в кожному Інтернет-магазині відіграють відгуки про товари магазину. Відгуки про товари забезпечують одразу кілька важливих функцій: по-перше, вони забезпечують канал зворотного зв'язку між покупцем та магазином; по-друге, відгуки дають змогу іншим користувачам дізнатися про досвід користування товаром та взаємодії з магазином від таких самих покупців.

Усе вищезгадане робить відгуки про товари Інтернет-магазинів надзвичайно важливою частиною електронної комерції.

2. Аналіз літературних даних та постановка проблеми

Проблемі аналізу природної мови присвячено достатньо багато уваги, що демонструється значною кількістю публікацій на дану тему.

Велика кількість досліджень, наприклад [1–4], підіймає питання аналізу методів машинного навчання та їх оптимізації. Ці роботи глибоко розкривають теоретичні засади різних типів машинного навчання в контексті обробки та аналізу природної мови, а саме навчання з учителем в різних його варіаціях [3], методи створення та оптимізації наборів навчальних даних для автоматичного навчання [1], неперервне автоматичне навчання на основі подальшого аналізу природної мови [4] та методи покращення результатів прогнозування [2]. Всі ці роботи та безліч інших висвітлюють широкий круг проблем та задач обробки природної мови, однак більшість з цих робіт присвячена загальній теорії аналізу текстів або специфічним частинам роботи та оптимізації алгоритмів та не беруть до уваги специфіку відгуків про товари магазинів електроніки, а саме додаткові атрибути відгуку, окрім самого тексту та ту особливість текстів відгуків, що вони зазвичай короткі, більшість відгуків з тестового набору має до двох речень.

Також, багато досліджень присвячено саме аналізу відгуків. Наприклад, в роботі [5] досліджується проблема зняття неоднозначності займенника ти/ви. Це дозволяє підвищити точність на наступних етапах моделювання. Однак, саме по собі визначення позначеного займенником об'єкту не має практичного застосування.

Базуючись на вищезгаданих роботах по темі аналізу природної мови та на інших, загальновідомих класичних дослідженнях, доцільним є проведення додаткових досліджень в напрямку аналізу відгуків на товари магазинів.

3. Мета та задачі дослідження

Метою даної статі є дослідження та покращення існуючих методів аналізу відгуків про товари Інтернет-магазинів електроніки та отримання інформації про досвід користування товаром або послугою та взаємодії з магазином, яка може бути використана для виявлення закономірностей та прогнозування властивостей відгуків, наприклад, прогнозування оцінки відгуку на основі його тексту.

Для досягнення мети були поставлені наступні задачі:

1. огляд існуючого стану розв'язання проблеми, виявлення протиріч відомих теоретичних або експериментальних результатів;

2. побудова математичних та комп'ютерних моделей для аналізу, класифікації та передбачення властивостей відгуків;

3. практична реалізація класифікатора на основі для кожної з моделей;

4. підготовка та очистка тестових даних, тренування моделей та оцінка отриманої якості класифікації та передбачення для кожної з моделей;

5. аналіз результатів класифікації та передбачення для кожної з моделей, порівняння результатів з очікуваними та їх інтерпретація в контексті задачі аналізу відгуків на товари магазинів електроніки.

4. Матеріали та методи дослідження

Сучасний Інтернет-магазин не можливо уявити без можливості залишення відгуків про товари та про сам магазин. Згідно досліджень, проведених у США та Канаді в 2014 році, 88 % покупців читають відгуки про магазини та товари перед прийняттям рішення про купівлю. Також у результаті дослідження було показано, що для користувачів однаково важливими є і кількість і якість відгуків [6].

В даній роботі розглянуто методи автоматизованого комп'ютерного аналізу відгуків про товари Інтернет-магазинів та прогнозування властивостей на основі аналізу тексту відгуку. Сучасний стан розвитку інформаційних технологій дає змогу автоматизувати аналіз відгуків, виявлення, не помітних на перший погляд, закономірностей та навіть прогнозування результату та якості відгуку.

Для досягнення цього були використані проаналізовані методи обробки природної мови на основі векторизації текстів відгуків за допомогою загальновідомого алгоритму TF-IDF та побудови прогнозуючих моделей.

5. Дослідження методів моделювання для відгуків на товари магазинів

В якості початкових даних була взята вибірка відгуків на мобільні телефони з Інтернет-магазину Amazon. Ця вибірка налічує більше 400 тисяч відгуків на різноманітні телефони від бюджетних до флагманських моделей різних модельних років.

Кожен відгук про товар магазину складається з декількох атрибутів:

– товар, про який залишено відгук;

– рейтингова оцінка. Задається клієнтом та приймає значення від 1 до 5;

– текст відгуку: довільний текст;

– корисність відгуку: кількість людей, які відмітили відгук як корисний.

Для наочності сприйняття в табл. 1 наведено декілька відгуків з вибірки з метою розуміння з чим ми маємо справу.

Сукупність цих атрибутів і робить предметну галузь відгуків на товари магазинів електроніки особливою та дає змогу будувати та навчати моделі використовувати комбінацію властивостей відгуку, а не лише тексту.

Дослідимо статистичні властивості вибірки. З цією метою на рис. 1 наведено графік розподілу кількості відгуків за довжиною. Як ми можемо переконатися, користувачі не мають схильності писати довгі відгуки на товари.

Таблиця 1

Частина вибірки, яка буде використана в якості вхідних даних

Назва	оцінка	Текст відгуку	Корисність
Apple iPhone 6	5	Excellent phone works great,Looks great ,Works fantastic	0
Apple iPhone 6	1	It was all chippy around the edges	6
Apple iPhone 6	4	over all good condition. work great	0

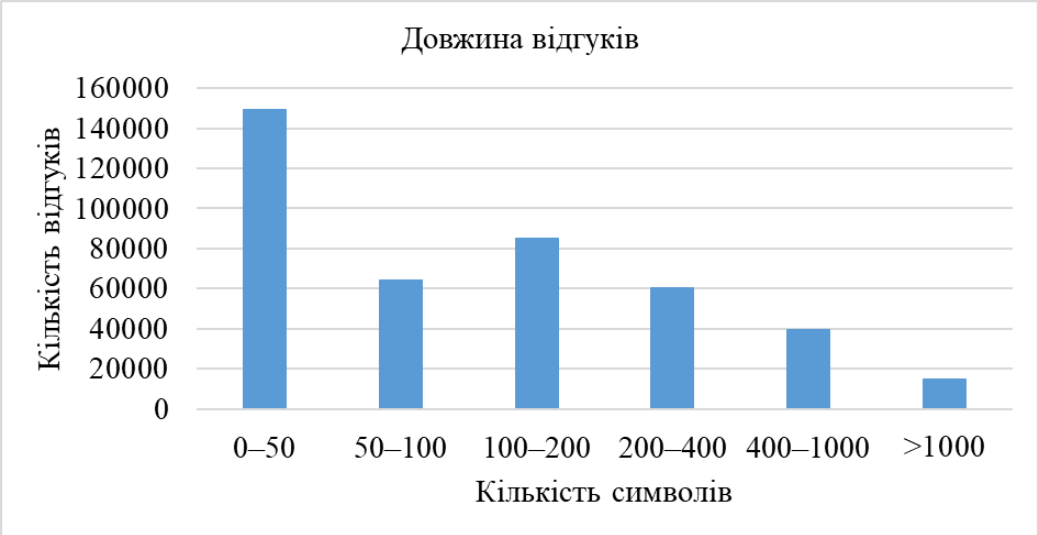

Рис. 1. Розподіл довжини тексту відгуків

В ході даної роботи буде проведено дослідження алгоритмів для обробки природної мови, якою викладений текст відгуку, на основі чого буде проведено прогнозування оцінки відгуку.

Нажаль, текст відгуку у чистому вигляді погано підходить для машинної обробки. Вони мають у своєму складі інформацію, яка не має прямого відношення до досліджуваних властивостей, а саме, різ-

ні символи, гіпертекстова розмітка. Також дані не є нормалізованими. Тому перед моделюванням нам необхідно підготовити дані для обробки. З цією метою зроблено наступні кроки:

– видалені небуквені символи;

– видалені стоп-слова.

Наступним кроком буде нормалізація даних. Важливим кроком для підвищення ефективності та точності результатів є приведення всіх даних до однієї форми. З цією метою зроблено наступні кроки:

– переведено всі слова в нижній регістр. Це допоможе об'єднати однакові слова незалежно від їх написання;

– скорочено всі слова до певної основи (стемінг). Цей процес називають злиттям. Цей крок дозволяє підвищити точність.

Для перевірки точності моделювання розділено вихідну вибірку на 2 частини у співвідношенні 3 до 7: більша частина буде використана для навчання моделей, а менша для перевірки результатів навчання моделей.

Після етапу підготовки даних текст відгуку позбувся шумових символів, загальних слів, які не несуть смислового навантаження (стоп-слів), дані нормалізовані, але все ще представлені в не придатному для обробки форматі.

Для обробки статистичними методами та методами глибинного навчання дані мають бути представлені у векторному вигляді. Для векторизації використано-алгоритм TFIDF. Трансформовано кожний відгук у вектор, де елемент визначає чи є певна ознака з набору у відгуку та у якій кількості. Загальний вигляд вектору ознак:

$$R_i = (w_1 \dots w_n), \quad (1)$$

де R_i – це вектор і-того відгуку, w_k – значення елементу вектору, яке приймає значення 0, якщо ознака відсутня у відгуку та числове значення, яке дорівнює кількості включень певної ознаки у відгуку.

При розбиванні тексту на ознаки було визнано доцільним використати n-грами, тобто абстракцію словосполучень. Дуже часто зустрічаються словосполучення, які не можливо проаналізувати за окремими словами [7]. В практичній реалізації були використані найбільш значущі n-грами від першого до четвертого порядку.

Після підготовки даних та проведення розбиття на ознаки і векторизації можна перейти до аналізу методів створення прогнозуючої моделі. Для порівняння різних методів була обрана задача прогнозування рейтингової оцінки товару користувачем на основі тексту відгуку, тобто ми намагаємося зрозуміти на скільки покупцю сподобався товар виходячи з його відгуку, що вимагає також і елементи аналізу тональності. Основним показником якості моделі обираємо точність прогнозування рейтингу товару на основі тексту відгуку.

В ході моделювання були порівняні такі математичні моделі:

– Наївна Баєсова класифікація. Ця класифікаційна модель заснована на використанні теореми Баєса. У загальному вигляді формула має такий вигляд:

$$classify(f_1 \dots f_n) = \\ = \arg max_c\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i \mid C = c) \quad (2)$$

– класифікатор на основі стохастичного градієнтного спуску;

– класифікатор на основі випадкового лісу (Random Forest). Його суть у використанні набору дерев ухвалень рішень. Фінальне рішення по класифікації приймається після проведення «голосування» кожного з дерев. Результатом є клас, за який «проголосувала» найбільша кількість дерев;

– класифікатор на основі градієнтного бустінгу. Цей метод заснований на комбінуванні простих моделей для побудови ефективної моделі;

– мультикласовий наївний Баєсов класифікатор на основі методу опорних векторів (*SVM with NB features (NBSVM)*)[8].

– глибинне навчання на основі довгої короткочасної пам'яті (архітектура рекурентних нейронних мереж, запропонована 1997 року Зеппом Хохрайтером та Юргеном Шмідгубером)[9] Перевагою цього методу є те, що цей тип нейронних мереж має змогу вивчати довгострокові залежності, що може бути дуже корисним при обробці тексту, який має певні залежності в реченнях, що можуть бути закешовані в пам'ять нейронної мережі;

– згорткові нейронні мережі [10]. Цей тип нейронних мереж традиційно використовується для задач обробки природної мови та показує хороші результати.

6. Результати моделювання та прогнозування

Для практичної реалізації порівнюваних методів була обрана мова програмування Python та бібліотека scikit-learn. Мова програмування Python широко використовується для наукового моделювання. Бібліотека scikit-learn – це просте та ефективне відкрите програмне забезпечення для аналізу даних, в тому числі й для аналізу природної мови.

В табл. 2 наведені результати практичної реалізації усіх вищезгаданих методів та апробації їх на наявному наборі тестових.

Таблиця 2

Результати точності передбачення рейтингової оцінки

Назва методу	Точність
Наївний Баєсов класифікатор	0.6569
Класифікатор на основі стохастичного градієнтного спуску	0.6818
Класифікатор на основі випадкового лісу	0.6565
Класифікатор на основі градієнтного бустінгу	0.6372
Мультикласовий наївний Баєсов класифікатор	0.6865
Класифікатор на основі довгої короткочасної пам'яті	0.6727
Класифікатор на основі згорткові нейронні мережі	0.6781

7. Обговорення результатів прогнозування для моделей

Відносно невисокий результат наївного Баєсового класифікатора є наслідком простоти моделі та припущень, які покладені в основу моделі. Такий висновок можна підтвердити тим, що найкращий результат показав покращений алгоритм наївного Баєсового класифікатору на основі методу опорних векторів.

Невисокий результат класифікаторів на основі випадкового лісу та градієнтного бустінгу зумовлений дуже високою розрідженістю даних, що є характерною рисою коротких заміток, таких, як відгуки на товари Інтернет магазинів. В перспективі можливе покращення точності роботи цих алгоритмів за допомогою введення додаткових обмежень зумовлених предметною галуззю, налаштування параметрів навчання та зменшенням розмірності векторизованих даних за допомогою відбирання в ознаки тільки найважливіші. Нажаль, останнє може негативно вплинути на максимально можливу точність передбачення.

Високий результат продемонстрували класифікатори на основі глибинного навчання нейронних мереж. Ці моделі очікувано проявили себе краще, ніж інші, що зумовлено їх достатньою складністю для обробки предметної галузі достатньо великої розмірності, якою і є аналіз відгуків на товари інтернет-магазинів. В перспективі можливе більше підвищення точності за допомогою збільшення розмірності векторизованих даних. Іншим методом підвищення точності може бути збільшення розмірності та глибини самих моделей нейронних мереж. Нажаль, обидва ці підходи негативно впливають на швидкодію алгоритму.

Неможливо однозначно відповісти на питання який з алгоритмів аналізу найкращий або найгірший. Кожен з них підходить для вирішення певного класу практичних задач та має потенціал вийти в лідери.

В якості подальших кроків в дослідженні методів аналізу відгуків на товари інтернет-магазинів можна розробити та дослідити властивості алгоритму, заснованому на комбінації декількох вищезгаданих моделей або дослідження більш специфічних обмежень для предметної галузі, що може виходити за рамки тільки лише однієї науки та знаходитися на стику психології та інформатики.

8. Висновки

В результаті виконання роботи були досліджені різні класи методів аналізу відгуків на товари магазинів електроніки, практиці порівняні їх можливості та результати передбачення. З цією метою були вирішені наступні задачі:

1. досліджено сучасний стан розв'язання проблеми аналізу відгуків на товари магазинів, в результаті чого було наочно висвітлено необхідність дослідження;

2. побудовано математичні та комп'ютерні моделі для аналізу відгуків на основі різноманітних класів методів (Баєсов класифікатор, класифікатор на основі випадкового лісу та інші);

3. реалізоване прикладне програмне застосування для практичної перевірки обраних математичних та комп'ютерних моделей;

4. в якості тестових даних була використана вибірка реальних відгуків з категорії розблокованих стільникових телефонів магазину Amazon. На ній були натреновані моделі та отримані результати прогнозування;

5. проаналізовано результати для кожної з моделей, порівняно з теоретично очікуваними результатами та інтерпретовано отримані результати в контексті задачі аналізу відгуків на товари магазинів електроніки.

Література

1. Data programming: Creating large training sets, quickly / Ratner A. et. al. // Advances in Neural Information Processing Systems (NIPS). New York: Curran Associates, 2016. P. 3567–3575.

2. Lei T., Barzilay R., Jaakkola T. Rationalizing neural predictions // Empirical Methods in Natural Language Processing. Austin, 2016. P. 107–117. doi: http://doi.org/10.18653/v1/d16-1011

3. Roth B., Klakow D. Combining generative and discriminative model scores for distant supervision // Empirical Methods in Natural Language Processing. Seattle, 2013. P. 24–29.

4. Srivastava S., Labutov I., Mitchell T. Joint concept learning and semantic parsing from natural language explanations // Empirical Methods in Natural Language Processing. Copenhagen, 2017. P. 1527–1536. doi: http://doi.org/10.18653/v1/d17-1161

5. Voigt R., Jurafsky D. The Users Who Say 'Ni': Audience Identification in Chinese-language Restaurant Reviews // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, 2015. P. 314–319. doi: http://doi.org/10.3115/v1/p15-2052

6. 88% Of Consumers Trust Online Reviews As Much As Personal Recommendations // Search Engine Land. URL: https://searchengineland.com/88-consumers-trust-online-reviews-much-personal-recommendations-195803 (Last accessed: 04.06.2018)

7. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank // Stanford University Sentiment Analysis. URL: https://nlp.stanford.edu/sentiment/ (Last accessed: 07.06.2018)

8. Wang S., Manning C. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju, 2012. P. 90–94.

9. Hochreiter S., Schmidhuber J. Long Short-Term Memory // Neural Computation. 1997. Vol. 9, Issue 8. P. 1735–1780. doi: http://doi.org/10.1162/neco.1997.9.8.1735

10. Backpropagation Applied to Handwritten Zip Code Recognition / LeCun Y. et. al. // Neural Computation. 1989. Vol. 1, Issue 4. P. 541–551. doi: http://doi.org/10.1162/neco.1989.1.4.541

Рекомендовано до публікації д-р техн. наук Шостак І. В.
Дата надходження рукопису 03.05.2018

Вечур Олександр Володимирович, кандидат технічних наук, доцент, кафедра програмної інженерії, Харківський національний університет радіоелектроніки, пр. Науки, 14, г. Харків, Україна, 61166
E-mail: avechur@gmail.com

Сподарець Олексій Михайлович, кафедра програмної інженерії, Харківський національний університет радіоелектроніки, пр. Науки, 14, г. Харків, Україна, 61166; E-mail: alexspodarets@gmail.com