ДОСЛІДЖЕННЯ МОДЕЛЕЙ ОЦІНЮВАННЯ ТОНАЛЬНОГО ЗАБАРВЛЕННЯ ТЕКСТУ

Литвинов М.Г., магістрант кафедри ПІ,
e-mail: maksym.lytvynov@nure.ua
Науковий керівник: д.т.н., проф. Єрохін А.Л.
Харківський національний університет радіоелектроніки

This paper considers the problem of automatic classification of estimation models of sentiment text color, approaches and methods to resolve it. The classification algorithms are proposed: the naive Bayes classifier and the method of k-nearest neighbors. The cross-checking method is used to evaluate the algorithm's performance. According to the results of the tests, the naive Bayes classifier with the help of a teacher has more chances for further development and improvement with more reliable accuracy of the obtained data.

Значний розвиток інформаційних технологій, зокрема в області обробки природної мови, змушує багатьох дослідників сфокусуватися на вивченні, пошуку нових та удосконалення вже існуючих шляхів роботи з великими обсягами текстових даних. Дана галузь не має єдиної технології визначення, адже вона перебуває у стані постійних досліджень та розробок. Однак, існують певні аспекти, які б об'єднували усі існуючі визначення[1]. Наш вітчизняний дослідник у галузі масових комунікацій, Олег Іванов, визначає аналіз природної мови як «міждисциплінарну галузь науки, що охоплює методики обчислювальної лінгвістики та теорії штучного інтелекту, основним проблемним полем якої є забезпечення взаємодії людських комунікативних актів (вербальних та невербальних) та комп'ютерних систем»[2].

Обробка природної мови належить до класу штучного інтелекту, до так званої AI-повної задачі (від англ. AI-complete), через його складність в розпізнаванні та в потребі величезних знань системи про навколишнє середовище. Ця задача потребує створення системи, котра буде розуміти та відповідати на природній мові так добре як людина. На сьогодення для більш точних результатів проблеми обробки не можуть бути розв'язані лише за допомогою сучасних комп'ютерних технологій, без використання людино-орієнтованих обчислень, тобто без передання частини завдань людині на певних етапах вирішення[3].

Наведемо на рисунку 1 графічно процес аналізу почуттів людини щодо будь-якого тексту:

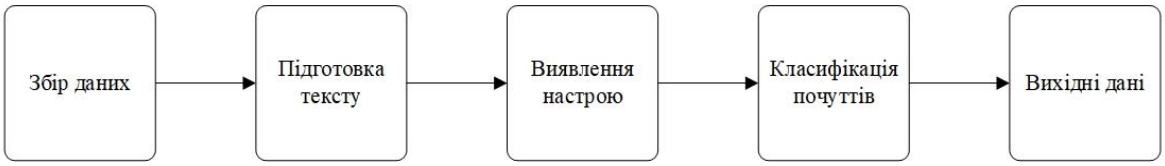

Рисунок 1 – Процес аналізу настрою тексту

На етапі збору даних джерелом може слугувати уся сучасна мережа інтернету: публічні форуми та блоги, відгуки користувачів різноманітних сервісів, огляди продуктів, приватні журнали, засоби масової інформації, такі соціальні мережі, як Facebook, Twitter, та ін.

Підготовка тексту передбачає очищення отриманих даних перед аналізом. Зазвичай вона виявляє та виключає нетекстовий вміст з текстового набору даних та будь-яку інформацію, яка не вважається релевантною для досліджуваної області. Таким чином видаляються зупинки слів або слів, які не мають відношення до курсу аналізу.

Виявлення настрою вимагає оцінювання та вилучення відгуків та думок з тексту за допомогою обчислювальних завдань. Кожне речення вивчається для суб'єктивності. Тільки судження з суб'єктивними виразами зберігаються в наборі даних. Висловлювання, які передають факти та об'єктивне спілкування, віддаляються від подальшого аналізу. Виявлення настроїв здійснюється на різних рівнях або окремим терміном, фразами, повними реченнями або повним документом із загальноприйнятими методами.

На етапі класифікації почуттів іде обробка кожного суб'єктивного речення чи слова в текстовому наборі даних у класифікаційні групи. Зазвичай ці групи представлені на двох крайніх точках континууму (позитивний, негативний, добрий, поганий, подібний, нелюбов). Однак, класифікація може також включати кілька точок, подібних до оцінок зірок, що використовуються, наприклад, готелями.

Загальна мета аналізу полягає в перетворенні неструктурованого фрагментованого тексту в змістовну інформацію. Після завершення аналізу використовується низка

звичайних параметрів для відображення результату текстового аналізу. Головним з них є використання графічних дисплеїв, таких як кругові діаграми, стовпчастих діаграм та лінійних графіків. Полярність сегментується за кольором, частотами, відсотками та розмірами. Формат подання залежить від інтересу дослідників.

Час може бути включений в аналіз. Як правило, це графічно відображається шляхом побудови лінії часу настрою шляхом побудови значень вибраної статистики (частоти прикладу, відсотків і середніх) за певний проміжок.

Можна виділити два основних підходи щодо визначення тонального забарвлення текстів: методи, засновані на правилах та словниках та методи машинного навчання.

Перший підхід для вирішення проблеми сентимент-аналізу заснований на правилах, базується на лінгвістиці. В ньому велику роль відіграє семантика слів і правила побудови речень. Такий підхід має на увазі наявність тонального словника, що містить слова або колокації. Для кожного слова в тональному словнику відзначена тональність і, іноді, сила тональності (наприклад, за шкалою від 1 до 10, де 10 – це сильна позитивна тональність). Тональний словник може бути взятий ззовні, або ж сформований статистично. Далі, відповідно до підходу, заснованого на правилах, з кожної пропозиції або його частини в первісному тексті рецензії формується синтаксичне дерево, що містить ланцюг із слів або колокацій, що залежать один від одного. Відбувається визначення об’єкта аналізу та напрямки тональності. Лінгвістичні правила, за якими формуються синтаксичні дерева, можуть бути сформовані відповідно до вподобань дослідника. Далі, певні слова або їх комбінації порівнюються з іншими із тональних словників і, таким чином, присвоюється напрямок тональності та, опціонально, сила тональності. Тональність всього тексту рецензії може формуватися на основі тональності його частин.

Другий підхід заснований на методах машинного навчання. В рамках цього підходу до вирішення проблеми сентимент-аналізу потрібен необхідний набір текстів в якості навчальної вибірки. Машинне навчання з учителем потребує набору заздалегідь розмічених текстів рецензій. Окремий екземпляр цього набору – пара із вектору ознак, який є поданням конкретного тексту, і тональності цього тексту. Тональність рецензії виявляється або самим автором, або експертом (наприклад, дослідником).

Під терміном «вектор ознак» мається на увазі векторна модель семантики. Вектор ознак – це уявлення кожного тексту рецензії як точки в багатовимірному просторі. Таке уявлення текстів рецензій у вигляді векторів ознак необхідно для подальшої класифікації, так як вектори можна порівнювати один з одним, шляхом обчислення відстаней між ними. Близько лежать один до одного вектори, які відповідають семантично схожим документам. Тобто векторне подання документу буде *j* $d_j = \{w_1, w_2, \dots, w_n\}$, де w_n – вага терміну *n* у документі *j, n* – кількість документів у вибірці. Поняття «терміна» може варіюватися - термін може відображати окреме або ключове слово, пару слів або навіть фразу з тексту. Кожен термін відноситься до окремого виміру, кількість параметрів вектору дорівнює *n*. Вага терміну вказує на його важливість в документі, тобто якщо термін з'являється в тексті, то він має ненульову вагу. Вага може бути обчислена різними способами.

В алгоритмах, заснованих на машинному навчанні з учителем, відбувається аналіз розміченого набору текстів (навчальної вибірки) і статистично формується патерн для використання при класифікації нових вхідних векторів. Машинне навчання без учителя ставить проблему виявлення шаблону на підставі нерозміченої вибірки. Через те, що тексти в вибірці нерозмічені, не існує засобів оцінки правильності роботи алгоритму (тобто правильності рішень, прийнятих класифікатором). Тому ймовірність помилки набагато більше, ніж в методах типу «навчання з учителем», і ефективність таких алгоритмів значно нижче.

Багато дослідників використовують комбінацію двох підходів: лінгвістичного та статистичного. Причина в тому, що змішаний підхід на практиці показує кращі результати[4, 5].

Лінгвістичний підхід дає досить гарні результати, бо дозволяє виявляти напрямок тональності для частин тексту, а не тільки для документу в цілому. Підхід, заснований на правилах, тісно пов'язаний з семантикою слів, і тому демонструє більш глибокий аналіз тексту. Але у цього підходу є і серйозні недоліки. Наприклад, лінгвістичний підхід сильно прив'язаний до мови, на якому написані тексти. Модель, розроблена для однієї мови, абсолютно непридатна для іншої, в першу чергу через наявність тональних словників, сформованих для конкретної мови, а також із-за синтаксису мови.

Наївний Байєсівський класифікатор (Naive Bayes Classifier, NBC) є одним із прикладів використання методів векторного аналізу. Дана модель класифікації базується на понятті умовної ймовірності приналежності документа *d* класу *c*.

NBC – один з найбільш часто використовуваних класифікаторів, через його простоту в імплементації та тестуванні. У той же час, наївний Байєсівський класифікатор демонструє не гірші результати, в порівнянні з іншими, більш складними класифікаторами.

В основі цього класифікатора лежить теорема (або формула) Байєса:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}.$$

Для даної моделі, документ – це вектор $d = \{w_1, w_2, \dots, w_n\}$, де w_i – вага -го терміну, а n – розмір словника вибірки. Таким чином, відповідно до теореми Байєса, ймовірність класу *c* для документу *d* буде:

$$P(c \mid d) = \frac{P(c)P(d \mid c)}{P(d)}.$$

Найбільш ймовірний клас c^*, якому належить документ d той, при якому умовна ймовірність приналежності документа *d* класу *c* максимальна:

$$c^* = arg\ max_c\ P(c \mid d).$$

За теоремою Байєса:

$$c^* = arg\ max_c\ P(d \mid c) * P(c)$$

та згідно з тим, що $d = \{w_1, w_2, \dots, w_n\}$, то

$$c^* = arg\ max_c\ P(w_1, w_2, \dots, w_n \mid c) * P(c).$$

Для наївного Байєсівського класифікатора визначено істотне припущення – передбачається, що всі ознаки документу *d* незалежні один від одного. Через це допущення модель і отримала назву «наївна». Це дуже серйозне спрощує допущення і, в загальному випадку, воно не так, але наївна Байєсова модель демонструє непогані результати, незважаючи на це [6, 7]. Передбачається також, що позиція терміну в реченні не важлива. Як наслідок, умовну ймовірність $P(w_1, w_2, \dots, w_n|c)$ для ознак можно визначити як:

$$P(w_1|c)(w_2|c) \dots (w_n|c) = \Pi_i P(w_i|c_i).$$

Необхідно оцінити $P(c_j)$ та $P(w_i|c_i)$. $P(c_j)$ являє собою відносну кількість документів класу *j* в навчальній вибірці до загальної кількості документів.

$$P(\mathrm{c}) = \frac{D_c}{D},$$

де D_c – кількість документів класу *c,* а *D* – загальна кількість документів у вибірці.

Для оцінки умовних ймовірностей для ознак, використовується формула:

$$P(w_i|c_i) = \frac{count(w_i, c_i)}{\sum w \epsilon V \, count(w_i, c_i)},$$

де $P(w_i|c_i)$ визначається як відношення кількості термінів w_i у класі c_i загальної кількості термінів у цьому класі, V – словник навчальної вибірки.

Метод *k* найближчих сусідів. Для його реалізації потрібна навчальна вибірка розмічених текстів. Для визначення класу тексту з тестової вибірки, потрібно визначити відстань від вектору цього тексту до векторів із навчальної вибірки. Визначити *k* об'єктів навчальної вибірки, відстань до яких мінімальна (*k* задається експертом або вибирається згідно з оцінками ефективності). Клас вхідного вектору – це клас, якому належать більше половини з сусідніх *k* векторів. В якості опції відстані було обрано Евклідову відстань:

$$p(x_i, x_j) = \sqrt{\sum_{i=1}^{n} (x_i - x_j)^2},$$

де $x_i = (x_i, \dots, x_i^n)$ – вектор *n* ознак *i*-го об'єкту, $x_j = (x_j, \dots, x_j^n)$ – вектор *n* ознак *j*-го об'єкту.

Якщо значення *k* буде маленьким, то може виявитися, що єдиним найближчим об'єктом буде об'єкт з неправильно певним класом, який дасть невірне рішення, такі випадки називають «викид». Якщо *k* матиме велике значення, то тоді «переможе» найпопулярніший клас. В такому випадку, відстань до об'єкта класифікації не має ролі. Компромісом вважають, коли $k = \sqrt{N}$.

В якості метрик правильності класифікації текстів були обрані точність і повнота. Точність в межах класу – це частка текстів, що дійсно належать даному класу, щодо всіх текстів, зарахованих класифікатором до цього класу. Повнота системи – відношення

числа знайдених класифікатором текстів, що належать класу, до числа всіх текстів цього класу в тестовій колекції.

Щодо класу позитивних текстів, точність і повноту будемо вимірювати наступним чином:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN},$$

де P та R – точність і повнота відповідно, TP – істинно-позитивне рішення, FP – помилково-позитивне рішення, FN – помилково-негативне рішення.

Для тестування алгоритмів визначення тонального окрасу текстів було взято відгуки на кіно-рецензії, 250 позитивних і 250 негативних відгуки. Для порівняння моделей був використаний метод перехресної перевірки.

Таблиця 1. Порівняння досліджуваних моделей

	Точність *P*, %	Повнота *R*, %
Наївний Байєсівський класифікатор	83,5	88,25
Метод *k* найближчих сусідів	61,7	70,5

В даній роботі проаналізовано моделі оцінювання тонального забарвлення текстів, виявлено основні підходи до вирішення цього питання.

Можна зробити висновок, що машина без сильного (майже людського) штучного інтелекту не має жодних додаткових знань чи вмінь для розв’язання тої чи іншої проблеми[8]. Але з допомогою учителя, штучний інтелект має всі шанси на коректне навчання будь-якої моделі сентиментального аналізу.

Краще розуміння людської емоції допоможе інтелектуальним технологіям створювати більше емпатичних програм та набувати досвіду в різноманітних галузях, керувати нашими вимогами, вдосконалювати методи навчання та з’ясовувати способи створення кращих продуктів, які відповідають нашим потребам.

Література.

1. Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc. — P.1.

2. Іванов О.В. Класичний контент-аналіз та аналіз тексту: термінологічні та методологічні відмінності / Іванов Олег Валерійович // Вісник Харківського національного університету імені В. Н. Каразіна, Харків: Видавничий центр ХНУ імені В. Н. Каразіна, 2013. — № 1045. — С.72.

3. Roman V. Yampolskiy. Turing Test as a Defining Feature of AI-Completeness . In Artificial Intelligence, Evolutionary Computation and Metaheuristics (AIECM) --In the footsteps of Alan Turing. Xin-She Yang (Ed.). pp. 3-17. (Chapter 1). Springer, London. 2013.

4. Karkaletsis, V., Klenner, M., Rentoumi, V., Petrakis, S., Vouros, G. United we stand: improving sentiment analysis by joining machine learning and rule based methods. // 7th International Conference on Language Resources and Evaluation (LREC 2010), Malta.

5. Prabowo, R. & Thelwall, M. Sentiment analysis: A combined approach. // Journal of Informetrics. – 2009. - № 3(2). – С. 143-157.

6. Pang L. Thumbs up? Sentiment Classification using Machine Learning Techniques // Proceedings of EMNLP (2002).

7. Domingos, P. & Pazzani, M. On the optimality of the simple Bayesian classifier under zero-one loss // Machine Learning. - 1997. - № 29. – С. 103-¬137.

8. Lenat, Douglas; Guha, R. V. (1989). Building Large Knowledge-Based Systems. Addison-Wesley. - P.1–5.