Тональний Аналіз як Напрямок Обробки Природної Мови

Владислав Біленький
кафедра програмної інженерії
Харківський національний
університет радіоелектроники
Харків, Україна
vladyslav.bilenkyi@nure.ua

Володимир Кобзєв
кафедра програмної інженерії
Харківський національний
університет радіоелектроники
Харків, Україна
volodymyr.kobziev@nure.ua

Дмитро Матвєєв
кафедра програмної інженерії
Харківський національний
університет радіоелектроники
Харків, Україна
dmytro.matvieiev@nure.ua

Tonal Analysis as a Direction of Natural Language Processing

Vladyslav Bilenky
department of Software Engineering
Kharkiv National University of
Radio Electronics
Kharkiv, Ukraine
vladyslav.bilenkyi@nure.ua

Volodymyr Kobziev
department of Software Engineering
Kharkiv National University of
Radio Electronics
Kharkiv, Ukraine
volodymyr.kobziev@nure.ua

Dmytro Matveev
department of Software Engineering
Kharkiv National University of
Radio Electronics
Kharkiv, Ukraine
dmytro.matvieiev@nure.ua

***Анотація*—Представлений сучасний стан тонального аналізу текстів та перспективи його розвитку.**

***Abstract*—The current state of texts tonal analysis and prospects of its development are presented.**

Ключові слова—природна мова, сентиментальний аналіз, тональність, текст, емоція.

Keywords—natural language, sentiment analysis, tonality, text, emotion.

I. Вступ

Обробка природної мови або комп'ютерна лінгвістика являє собою окремий напрямок штучного інтелекту (AI) і математичної лінгвістики. Вона дозволяє витягувати різноманітну корисну інформацію з тексту на природній мові. Одним із перспективних напрямків комп'ютерної лінгвістики є аналіз тональності тексту.

Розуміння емоцій – це один з найважливіших аспектів особистого розвитку та зростання, і, таким чином, це ключова плитка для емуляції людського інтелекту. Окрім важливості для просування AI, обробка емоцій також важлива для тісно пов'язаного завдання визначення полярності. Фактично, можливість автоматично зафіксувати настрої широкої громадськості щодо соціальних подій, політичних рухів, маркетингових кампаній та переваг товару викликала все більший інтерес як у науковому співтоваристві, так і до хвилюючих відкритих викликів, і в діловому світі, за чудові наслідки маркетингу та прогнозування фінансового ринку. Це призвело до появи сфер впливу афективного обчислення та аналізу настроїв, що впливає на взаємодію між людьми та комп'ютером, пошук інформації та обробку мультимодальних сигналів для вилучення настроїв людей з постійно зростаючої кількості соціальних даних в Інтернеті.

На стику такого роду наукових і бізнес-інтересів з'явилися два нових напрямки досліджень: аналіз думок, тобто їх витяг, і аналіз тональності (сентиментальний аналіз). Варто зазначити, що аналіз тональності прийнято вважати підзадачею з області аналізу думок. У даних областях активно використовуються методи інтелектуального аналізу даних (data mining) і обробки текстів на природній мові (natural language processing).

Аналіз тональності тексту [1] – клас методів у комп'ютерній лінгвістиці, призначених для автоматизованого виявлення в текстах емоційно забарвленої лексики і емоційної оцінки думок авторів по відношенню до об'єктів, мова про які йде в тексті. Більшість сучасних систем використовують бінарну оцінку, тобто «позитивний» або «негативний» сентимент, деякі системи здатні виявляти силу (глибину) тональності.

Дослідження завдання аналізу тональності текстових даних почалися відносно нещодавно та точного ідеального рішення на даний момент не існує, існують різні варіації відносно кожної специфіки моделі.

Вітчизняних робіт у даній галузі дуже мало. Англомовних робіт достатньо багато у виданнях усього світу. Науковці звідусіль спочатку шукають праці з теми тонального аналізу у міжнародних наукометричних базах даних, а потім вже вивчають і удосконалюють досліджені моделі на прикладі своєї мови. Звісно кожна мова має свою специфіку за рахунок різниці структури, писемності, але здебільшого можна сказати, що різниця між мовами і діалектами однієї мови умовна.

II. Основна частина

Протягом останніх років значні успіхи були досягнуті у здатності систем штучного інтелекту (AI) розпізнавати та аналізувати людські емоції та настрої, в значній мірі завдяки прискоренню доступу до даних, недорогих обчислювальних потужностей і розвинення глибоких можливостей навчання разом з обробкою природних мов (NLP) та комп'ютерним баченням [2].

Сентиментальний аналіз як напрямок обробки природної мови відноситься до так названих AI-повних задач наголошуючи на тому, що обчислювальна складність цих задач еквівалентна складності вирішення головного завдання штучного інтелекту — створення комп'ютерів, настільки ж розумних, як і люди. Задача, котру називають AI-повною, вважається такою, що не може бути розв'язаною за допомогою простого алгоритму. Такі задачі не можуть бути розв'язані лише за допомогою сучасних комп'ютерних технологій без використання людино-орієнтованих обчислень. Ця властивість може бути корисною, наприклад, для перевірки присутності людини за допомогою тесту, а також в галузі комп'ютерної безпеки для запобігання атакам методом "грубої сили", тобто простим перебором наявних варіантів.

Згідно з новим звітом Tractica, ці тенденції починають стимулювати зростання ринку програмного забезпечення для аналізу настроїв та емоцій. Tractica прогнозує, що до 2022 року світовий дохід від програмного забезпечення для аналізу настроїв та емоцій збільшиться до 237,3 мільярдів доларів. Група компаній з розвідки ринку прогнозує, що це зростання буде обумовлено рядом ключових галузей, включаючи роздрібну торгівлю, рекламу, бізнес-послуги, охорону здоров'я та ігровий процес (див. рис. 1).

Відповідно до результатів аналізу, найважливіші категорії випадків для визначення настроїв та емоцій будуть такими: клієнтський досвід, маркетингові дослідження; обслуговування клієнтів, охорона здоров'я, ігри, освіта та автомобільна промисловість.

Краще розуміння людської емоції допоможе інтелектуальним технологіям створювати більше емпатичних клієнтів та досвід в галузі охорони здоров'я, керувати нашими вимогами, вдосконалити методи навчання та з'ясувати способи створення кращих продуктів, які відповідають нашим потребам.

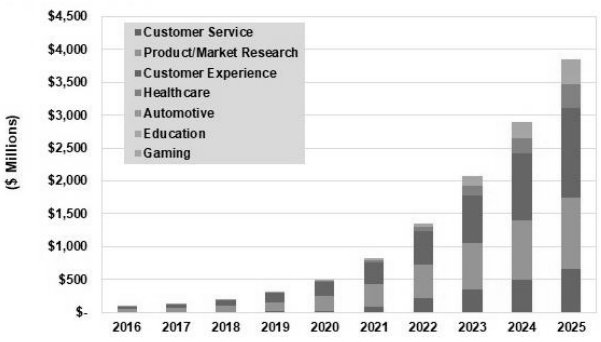

Рис. 1 – Прибуток ринку від ПЗ емоційного аналізу

Існуючі підходи до аналізу афективних обчислень та настроїв можна згрупувати у три основні категорії: технології, основані на знаннях, статистичні методи та гібридні підходи.

Текст розбивається на категорії, що залежать від наявності досить однозначних схожих слів, наприклад "щасливий", "сумний". Основною слабкістю підходів, що базуються на знаннях, є погане визнання впливу, коли беруть участь мовні правила. Наприклад, хоча база знань може правильно класифікувати речення «сьогодні був щасливим днем» як щасливу, воно, ймовірно, не зможе виконати речення «сьогодні не був щасливий день взагалі». З цією метою більш складні підходи, що базуються на знаннях, використовують правила лінгвістики, щоб розрізняти, як кожен конкретний запис бази знань використовується в тексті. Більш того, дійсність підходів, що базуються на знаннях, значною мірою залежить від глибини та широти зайнятих ресурсів. Інше обмеження підходів, що ґрунтуються на знаннях, полягає в типовій формі представлення їх знань, яка зазвичай суворо визначена і не дозволяє керувати різними нюансами концепції, оскільки висновок семантичних та афективних рис, пов'язаних з поняттями, обмежений фіксованим, плоским представленням.

Перехід від простого читання до прийняття участі у оцінюванні веб-сайтів змушує користувачів з більшим ентузіазмом ділитися своїми емоціями та думками через соціальні мережі, онлайн-спільноти, блоги, форуми та інші спільні засоби масової інформації. Останнім часом цей колективний розвідник поширився на багато різних напрямків Інтернету з особливим фокусом на областях, пов'язаних з нашим повсякденним життям, такими як торгівля, туризм, освіта, преса та здоров'я [3].

Проте, незважаючи на значний прогрес аналіз емоційних обчислень та настроїв все ще знаходить власний голос як нова міждисциплінарна галузь. Інженери та комп'ютерні вчені використовують методи машинного навчання для автоматичної класифікації впливу від відео, голосу, тексту та фізіології.

Поки що підхід до виявлення настроїв з тексту або мови базується, головним чином, на моделі "мішок слів", оскільки, на перший погляд, найпершою одиницею мовної

структури є слово. Однак вирази з одним словом – це лише підмножина понять, у той час коли вирази з декількох слів несуть певну семантику, яка зазвичай пов'язана з об'єктами, діями, подіями та людьми. Семантика, зокрема, вказує афективну (емоційну) інформацію, пов'язану з реальними сутностями, що є ключовим для розпізнавання емоцій та виявлення полярності, основними завданнями афективного обчислення та аналізу настроїв [4].

Гібридні підходи спрямовані на краще розуміння концептуальних правил, які регулюють настрої та ключі, які можуть передати ці поняття від реалізації до вербалізації в людському розумі. В останні роки подібні підходи поступово включають аналіз афективного обчислення та настроїв як міждисциплінарні поля між простою NLP та розумінням природної мови шляхом поступового переходу від синтаксичних методів до все більш і більш смислової системи, де обидва концептуальні знання та структура речення враховуються (див. рис. 2).

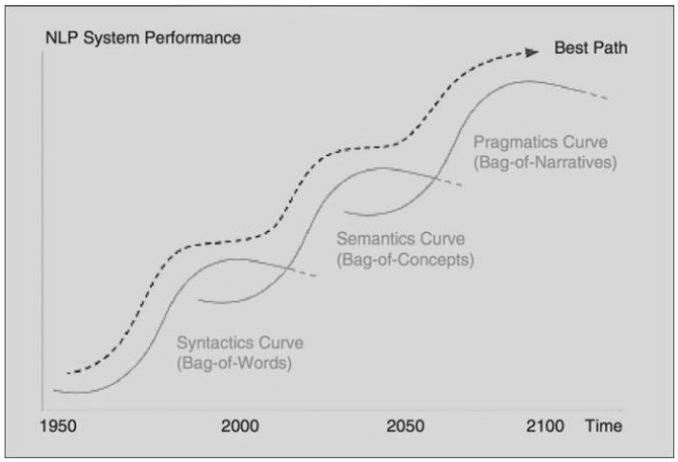

Рис. 2 – Розвиток продуктивності системи NLP

Зокрема, комбіноване застосування лінгвістики та баз знань дозволить відчути потік від концепції до поняття, заснованого на співвідношенні залежностей вхідної пропозиції, тоді як машинне навчання буде виконувати роль резервного копіювання відсутніх концепцій та невідомих лінгвістичних моделей.

Сентиментальний аналіз, як одна із AI-повних задач дозволяє витягувати різноманітну інформацію, що знаходиться у формі тексту на природній мові, а саме:

– маркетингові дослідження (вивчення споживчих переваг, вимір ступеня задоволення потреб споживачів, визначення ефективності поширення продуктів або послуг);

– фінансові сектор (пошук позитивної і негативної інформації щодо працівників або кліентів банку, акційних фондів, тощо);

– політологічні дослідження (збираються дані про політичні погляди населення. Це може мати істотне значення для кандидатів, які виступають від різних партій. Такий підхід застосовується організаторами передвиборної кампанії для виявлення того, що думають виборці по відношенню до різних проблем, і як вони пов'язують ці проблеми зі словами і діями кандидатів);

– рекомендації (аналіз відгуків та оглядів різних продуктів з метою допомоги покупцям при виборі товару, система не буде рекомендувати продукт або його постачальника при наявності негативних відгуків);

– аналіз новинних ресурсів щодо різних персон і подій;

– соціологічні дослідження (аналізуються дані з соціальних мереж, інших медіа);

– аналіз зворотного зв'язку від користувачів (при діалозі з користувачем система розпізнає його емоції і в разі незадоволення перемикає зв'язок на оператора);

– освіта (пошук кращих вищих навчальний закладів, підготовчих курсів, шкіл).

Висновки

Питання застосування систем обробки природної мови на основі методів та алгоритмів семантичного аналізу багато досліджують протягом останніх десяти років. На сьогоднішній день існує проблема досить чіткого знаходження емоційного стану текстових даних через швидкий доступ до інформації, взаємодії з метою вирішення питань споживачів, комунікації з клієнтами, проведення аналітичних досліджень, збору необхідної інформації, підвищення сервісу та інших переваг.

Системи видобутку настроїв наступного покоління потребують ширшої та більш глибокої загальноприйнятої бази знань, а також більш мозкових і психологічно мотивованих методів обґрунтування, щоб краще зрозуміти природні мовні думки і, отже, ефективніше подолати розрив між (неструктурованими) мультимодальними даними та (структурованими) машинообробленими даними.

У майбутньому поєднання наукових теорій емоцій з практичними інженерними цілями аналізу настроїв природної мови та поведінки людини прокладе шлях до розробки більш біологічно наближених підходів створення інтелектуальних пошукових систем, здатних застосовувати семантичні знання, робити аналогії, вивчаючи нові афективні знання та виявляючи, сприймаючи та "відчуваючи" емоції.

Література References

[1] Liddy, E.D. Natural Language Processing. In Encyclopedia of Library and Information Science. - NY. Marcel Decker Inc., 2001.— pp. 2-15.

[2] Іванов О.В. Класичний контент-аналіз та аналіз тексту: термінологічні та методологічні відмінності / Вісник Харківського національного університету ім. В.Н. Каразіна, - Харків: Видавничий центр ХНУ ім. В.Н. Каразіна. – № 1045, 2013, с. 69-73.

[3] Lenat D.B., Guha R.V. Building Large Knowledge-Based Systems. Addison-Wesley, 1989. – pp. 21-36.

[4] Artstein, R., & Poesio, M. Inter-coder agreement for computational linguistics. Journal of Computational Linguistics, 34(4), 2008. - pp. 555-596.