УДК 681.3

МЕТОДИ АНАЛІЗУ ТОНАЛЬНОСТІ ТЕКСТУ В СИСТЕМАХ АНАЛІЗУ ОПІНІЙ КОРИСТУВАЧІВ СОЦІАЛЬНИХ МЕРЕЖ

Кульчицький А.Б., Турко Ю.В.
Західноукраїнський національний університет, магістранти

I. Постановка проблеми

На сьогоднішній день в мережі Інтернет є велика кількість ресурсів, призначених для вираження думки людей про різні товари і послуги. Ці думки важливі як для самих користувачів при прийнятті рішень з приводу товару, так і для виробників при відстеженні споживчої якості своєї продукції. У 2016 році команда маркетингу та аналізу ринку компанії з надання послуг SaaS – BrightLocal1 провела опитування, в результаті якого було з'ясовано, що 88% респондентів вважають, що читання позитивних або негативних відгуків в Інтернеті впливає на їх рішення при придбанні товарів або послуг. Звідси випливає, що майже 9 з 10 споживачів роблять висновок про якість сервісу/товару на основі клієнтських відгуків [1].

У зв'язку з цим, виникає потреба обробляти великі обсяги інформації для визначення ставлення користувачів до того чи іншого об'єкту торгівлі. Щодня в мережі Інтернет публікується величезна кількість відгуків, що робить «ручний» збір та аналіз інформації в даному випадку непридатна. З цієї причини широкого поширення набула область комп'ютерної лінгвістики, спрямована на автоматичну обробку текстів на природній мові.

Завдання обробки текстів можна розбити на дві умовні категорії [2]. До першої відносяться завдання, з якими щодня стикається будь-який користувач: перевірка орфографії, фільтрація спаму. З точки зору дослідників в галузі автоматичної обробки текстів (АОТ), всі ці завдання майже вирішені, і сьогодні більш актуальні завдання з другої категорії, які потребують обробки великих текстових масивів: аналіз думок і відгуків, знаходження релевантних відповідей на питання (завдання «питання-відповідь»), конструювання рекомендаційних систем, що працюють з великими масивами неструктурованих даних. Відмітна особливість таких завдань – їх складність і відсутність формалізації, що призводять до того, що для них поки що немає повноцінного набору рішень, а застосовуються допоміжні методи класифікації текстів і виділення ключових слів і словосполучень .

Таким чином, розробка систем аналізу думок споживачів в соціальних мережах є актуальним завданням на сьогоднішній день. Такі системи корисні як покупцям, надаючи додаткову інформацію про характеристики товарів, так і компаніям, здійснюючи оперативний збір та аналіз думок для своєчасного реагування компаній на відгук споживача, коригування стратегії і оптимізації власного бізнесу.

II. Мета роботи

Метою роботи є розробка системи, що дозволяє в режимі реального часу проводити автоматичний збір і аналіз думки споживачів в соціальних мережах. Досягнення поставленої мети передбачає вирішення таких завдань: аналіз предметної області; вивчення існуючих методів виділення аспектів і аналізу тональності тексту; вивчення існуючих систем аналізу тональності думок споживачів в соціальних мережах; проектування, розроблення та тестування системи. В праці проведено аналіз існуючих методів виділення аспектів і аналізу тональності тексту.

III. Методи аналізу тональності тексту

Під аналізом думок споживачів в соціальних мережах мається на увазі добування думок з текстів з подальшою класифікацією текстів на основі тональності на три (позитивний, негативний, нейтральний) класи.

Тональність тексту – емоційна оцінка, виражена автором, щодо деякого об'єкта [3].

В аналізі тональності тексту вважається, що текстова інформація в мережі Інтернет ділиться на два класи: факти і думки. Ключовим поняттям є визначення думки, яка буває двох типів «просте думка» або «порівняння» [3].

Проста думка. Проста думка містить точку зору автора про один об'єкт. Проста думка може бути висловлено прямо: «Якість звуку динаміків була чудовою» або неявно: «Після придбання зволожувача повітря, в будинку стало набагато комфортніше дихати».

В аналізі тональності тексту для думки першого типу дається формальне визначення:

Простою думкою називається кортеж з п'яти елементів:

{{Entity}, {feature}, {sentiment value}, {holder}, {time}}, (1)

де {entity} – об'єкт, {feature} – аспект об'єкта, {holder} – автор думки, {sentiment value} – емоційна оцінка, {time} – момент часу.

Виділяються 3 види емоційної оцінки (sentiment value): позитивна;негативна;нейтральна.

Під нейтральною мається на увазі, що текст не містить емоційну складову по відношенню до об'єкта або в принципі. Об'єктом є людина, організація, подія, товар або тема обговорення. Аспект має на увазі під собою компоненти об'єкта і атрибути компонентів. Окремим випадком аспекту є сам об'єкт [4].

Порівняння. В аналізі тональності тексту для думки другого типу дається формальне визначення:

Порівнянням називається кортеж з шести елементів:

{{E1},{E2}, {A}, {po}, {holder}, {time}}, (2)

де E1 і E2 – множини порівнюваних об'єктів, A – аспект об'єктів, po – множина об'єктів, які вважав за краще автор (holder), time – момент часу публікації думки.

На відміну від кортежу, що визначає думку першого типу, кортеж думки другого типу не містить оцінку емоцій автора.

Другий тип думок можна розділити на три види [5]:

1) порівняння аспектів об'єктів на користь одного (Non-equalGradable);

2) прирівнювання аспектів різних об'єктів (Equative);

3) перевагу одного об'єкта над іншими (Superlative).

Порівняння першого типу мають вигляд «аспект об'єкта 1 перевершує в чомусь аспект об'єкта 2», наприклад, «Якість збірки смартфонівiPhone набагато краща, ніж у SAMSUNG». Другий тип виражає схожість аспектів різних об'єктів. Наприклад, «I Android, і iOS однаково зручні для розробки додатків під них». Прикладом третього типу, існує речення «СмартфонSamsung S6 Edge виявився найкращим з представлених на ринку».

Суб'єктивність. В аналізі тональності тексту часто зустрічається термін, пов'язаний з поняттям думки – суб'єктивність. У своїй роботі Пенг і Лі дають таке визначення об'єктивного і суб'єктивного речень. Об'єктивне речення висловлює фактичну інформацію про будь-що, тоді як суб'єктивне речення висловлює чиїсь особисті почуття і припущення.

Речення, що містять думку, зазвичай є суб'єктивними, тому аналіз тексту на наявність суб'єктивної інформації часто є підзадачею визначення полярності тексту.

Під визначенням полярності тексту мається на увазі класифікація текстів. Класифікація текстів – одне із завдань інформаційного пошуку, яке полягає у віднесенні тексту до однієї з декількох категорій на підставі змісту тексту. Це може бути:

- бінарна («позитивний» – «негативний»),
- тернарна («позитивний» – «негативний» – «негативний»),
- n-арна класифікація (з введенням таких проміжних класів тональності, як «помірно позитивний», «сильно негативний» і ін.).

Формальна постановка задачі визначення полярності має такий вигляд.Нехай X – множина текстових фрагментів, Y – кінцева множина класів тональності, φ: X → Y – цільова функція тональності, значення якої відомі тільки на кінцевій підмножині повідомлень навчальної вибірки:

$X_{train} = \{(x1, y1),..., (xn, yn)\}$.

Потрібно визначити апроксимуючу функцію тональності φ: X → Y, здатну класифікувати довільні текстові фрагменти. Функцію φ називають класифікатором. Класифікатор – це алгоритм, який співвідносить якісь вхідні дані з одним або декількома класами.

Висновки

У роботі було проведено дослідження та аналіз методів виділення аспектів і аналізу тональності тексту. В подальшому дослідженні було проведено аналіз ефективності методів для вибору оптимального класифікатора для розглянутої задачі.

Список використаних джерел

1. Frantzi K., Ananiadou S. Automatic recognitionof multi-wordterms. International Journal of Digital Libraries 3, 2000. – P. 117-132.
2. Larose D.T.Discovering knowledge in data: an introduction to data mining. – NewJersey: JohnWiley&Sons, Inc., 2005. – 240 p.
3. Lerman J. Programming Entity Framework. – CA: O'ReillyMedia, 2010. – 920 p.
4. Liu B. Sentiment Analysisand Subjectivity. [Електронний ресурс] – Режим доступу до ресурсу: https://www.cs.uic.edu
5. Mining and summarizing customer views. [Електронний ресурс] – Режим доступу до ресурсу:: https://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf