

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ТА ПРОГРАМНА СИСТЕМА ОЦІНЮВАННЯ ОПІНІЇ ВИСЛОВЛЮВАНЬ

© Лупенко С.А., Хомів Б.А., 2014

Запропоновано інформаційну технологію, котра основана на використанні лінгвістичних змінних та застосуванні вагових коефіцієнтів оцінювання емоційного забарвлення висловлювань щодо певного об'єкта. Реалізована програмна система дає змогу в візуальному інтерфейсі відстежувати кількісні показники опінії щодо об'єктів, та порівняти їх інтегральні показники із урахуванням вагових коефіцієнтів компонент.

Ключові слова: опінія, емоційне забарвлення, оцінювання опінії, інформаційна технологія, об'єкт, інтегральний показник, лінгвістична змінна, ваговий коефіцієнт.

Article is devoted to the information technology and software for opinion mining and sentiment analysis of messages. The information technology based on using linguistic variables and applying the weighting coefficients for sentiment analysis concerning a certain object is proposed in the article. Implemented software allows tracking quantitative characteristics of opinion on objects and comparing their integral indicators in visual interface.

Key words: opinion, sentiment, sentiment analisys, object, integral indicator, linguistic variable, weighting coefficient.

Вступ

Важливим завданням сьогодення є розроблення та впровадження інформаційних технологій у різні сфери діяльності людини. Не менш важливим є застосування сучасних ІТ розробок у галузі opinion mining – виявлення та оцінювання позитивності чи негативності висловлювань стосовно певного об'єкта дослідження.

Розроблення новітніх інформаційних технологій у галузі opinion mining зводиться до таких основних завдань:

- добування опінії – процес знаходження у текстових даних позитивних та негативних висловлювань;
- оцінювання опінії – процес присвоєння позитивним чи негативним висловлюванням певного числового еквівалента;
- узагальнення опінії – процес узагальнення позитивних та негативних висловлювань до певного інтегрального показника з метою порівняння об'єктів дослідження.

Під опінією мається на увазі позитивна чи негативна думка або судження про певний досліджуваний об'єкт, ним можуть бути компанії, бренди, люди, події.

Безсумнівна важливість реалізації таких інформаційних технологій для бізнесу загалом і для певних галузей, прибуток в яких багато в чому опосередкований характером медіа-присутності. У контексті цього ключової роль відіграють моніторингові платформи, конкурентоспроможність яких залежить від впровадження революційних інформаційних технологій, що дають змогу оперативно надавати актуальну інформацію клієнту.

Аналіз останніх досліджень та публікацій. Постановка задачі

Враховуючи широкий спектр сучасних систем оцінювання опінії, опишемо їхній розвиток з погляду точності та якості класифікації. Найпершими та найпримітивнішими підходами до оцінювання опінії висловлювань вважаються методи машинного навчання.

Методи машинного навчання

Методи машинного навчання ґрунтуються на автоматичній класифікації текстів, зарахуванні до певного класу (позитив чи негатив), як правило, з урахуванням кількості входжень слів. Детально машинне навчання описано у [1, 2] та інших джерелах.

Для демонстрації роботи методів машинного навчання автор реалізує прототип системи, що базується на методі Байеса [3], – найвійний байесівський класифікатор. Цей метод найпростіший в реалізації і за його допомогою легко продемонструвати суть машинного навчання. Загальний алгоритм складається з таких етапів:

1. Побудова навчального матеріалу. Це, як правило, набори висловлювань, ап'яріорі поділені на два класи: позитивний та негативний.

2. Навчання класифікатора на складених навчальних наборах. В цьому випадку – це метод Байеса, хоча можуть бути й інші, наприклад, метод опорних векторів, який, між іншим, показує, хоч і незначно, але кращі результати.

3. Класифікація висловлювань. Класифікація відбувається відповідно до вибраного методу. У методах машинного навчання основною мірою класифікації є кількість входжень слова у навчальні набори.

4. Оцінка повноти та точності класифікації. Оцінка повноти та точності класифікації визначається за допомогою практичного застосування класифікатора на навчальних наборах. Наприклад, беруть другу частину від ап'яріорі позитивних висловлювань та пропускають через класифікатор. Відповідно до цього і знаходять точність.

Варто звернути увагу, що якщо на вхід подається слово, якого немає в навчальних наборах, – класифікатор не знає, як поводитись в такій ситуації, тому ці методи і вважаються найпростішими, хоч і точність класифікації в деяких випадках наближається до 0.9.

Зрозуміло, що методи машинного навчання, хоч і дають порівняно високу точність оцінювання опінії, проте ніяк не враховують позитивних чи негативних конструкцій мови. Для отримання достовірнішого результату необхідно використовувати інші підходи, такі як використання словників наборів та шаблонів речень, котрі б дали змогу виявляти лише потрібні конструкції побудови мови.

Використання словників наборів та шаблонів речень

Цей підхід передбачає використання наборів позитивних та негативних словників. Крім того, використовують шаблони речень для видобування опінії. У роботах [4, 5] описано декілька шаблонів можливих словосполучень у реченні. Загальний алгоритм оцінювання опінії за допомогою цього методу полягає у такому:

1. Накопичення слів для словників позитивних та негативних слів. Такими словами можуть бути: хороший, поганий, гарний, нормальній, красивий тощо.

2. Написання шаблонів речень для ідентифікації входжень певних типів словосполучень у речення. При цьому враховуються частини мови певних слів у реченні. Приклади шаблонів:

іменник + прикметник	фотоапарат хороший
іменник + дієслово + прислівник	фотоапарат працює гарно

3. Аналіз висловлювань. На цьому етапі відбувається процедура розподілу речення на частини мови, так званий Part-of-Speech (POS)-парсер, після чого до речення застосовуються шаблони.

4. Оцінювання опінії. Якщо висловлювання підходить під певний шаблон і в ньому є слово, що міститься у словниках, – речення класифікується як позитивне чи негативне відповідно.

Для програмної реалізації цього підходу автор використовує POS-парсер [6] та словники [7] з ап'яріорі позитивними та негативними словами.

Варто зауважити, що на якість та точність добування опінії цим методом впливають такі два чинники: кількість шаблонів висловлювань та розмір словників із позитивними та негативними словами.

Отже, для аналізу висловлювань описаним вище методом необхідно мати досить великі словники позитивних та негативних слів і скінченну кількість шаблонів, котрі можуть описувати висловлювання.

Використання словників наборів та шаблонів речень дає кращі результати добування опінії в плані осмислення оцінюваного результату. Якщо порівнювати із методами машинного навчання, то текст може бути повною нісенітницею – а результат все одно залежатиме від кількості

входжень слів. Проте у разі використання шаблонів речень аналізуватимуться лише висловлювання зі змістом, котрі відповідатимуть шаблону.

Проте наскільки б великими ці словники не були, навіть якщо є безліч шаблонів висловлювань, ми не зможемо досягти такої точності видобування опінії, як із застосуванням семантичного аналізу речень.

Семантичні зв'язки та розуміння сенсу речень

Семантичний аналіз речень передбачає побудову семантичного дерева речення із повною взаємозалежністю слів у ньому.

Прикладами такого аналізу можуть бути системи Dictum [8], Lexalytics [9] або ABBYY Compreno [10]. Варто зазначити, що компанія SemanticForce [11] використовує саме такий підхід.

Безперечно, використання цього підходу, як і використання шаблонів речень, для видобування опінії потребує словників ізaprіорі позитивними та негативними словами. Проте, на відміну від використання шаблонів висловлювань, у разі графоморфологічного розбору речення не втрачаються зв'язки між словами, відповідно це впливає на точність видобування опінії, що, своєю чергою, вливає на якість оцінювання загалом.

Таблиця 1

Методи добування опінії

Методи	Характеристики			
	Простота реалізації	Потреба у POS аналізі	Потреба словників	Розуміння сенсу речень
Машинне навчання	+	-	-	-
Шаблони речень	+	+	+	-
Графоморфологічний аналіз	-	+	+	+

В кінцевому варіанті якість класифікації висловлювань на позитивність чи негативність залежить від розуміння семантики речень та якості словників.

Формулювання цілей статті

Мета

Розробити інформаційну технологію та створити програмну систему оцінювання опінії висловлювань.

Завдання

Для досягнення поставленої мети статті необхідно вирішити такі завдання:

1. Розробити інформаційну технологію оцінювання опінії висловлювань.
2. Створити програмну систему оцінювання опінії висловлювань.

Основний матеріал

1. Інформаційна технологія оцінювання опінії

Зрозумівши, як відбувається добування опінії, потрібно зрозуміти, як її оцінюють. Адже висловлювання “це хороша фотокамера” та “це неймовірно хороша фотокамера” мають різну важливість та вагу. Звичайно, вони є позитивними, проте з різним ступенем позитивності.

Для розуміння відмінності в оцінюванні між наведеними вище висловлюваннями автори запропонували інформаційну технологію оцінювання опінії (рис. 1). Інформаційна технологія передбачає добування, аналіз та оцінювання опінії із подальшим підрахунком до інтегрального показника з метою порівняння об’єктів.

Інформаційна технологія поділена на етапи (згідно із структурою, зображену на рис. 1), кожен з етапів характеризується методами та засобами реалізації певних методів.

Аналіз висловлювань користувачів описано у огляді існуючих рішень. Варто звернути увагу на табл. 1, у ній наведено методи та їхні характеристики, графоморфологічний аналіз речень є найактуальнішим нині. Після аналізу речення відбувається пошук об’єкта, його компонент та власне самої опінії, після чого здійснюється побудова графу об’єкта, відповідно до [1, 12].

Інформаційна технологія оцінювання опінії

Методи

Засоби



Рис. 1. Структурна схема інформаційної технології оцінювання опінії

1.1. Застосування лінгвістичних змінних

Детально застосування лінгвістичних змінних автори описали у роботі [12]. Такий підхід надає низку переваг. Якщо раніше, оцінюючи позитивне висловлювання, отримували “+1”, а негативне висловлювання – “-1”, то автори пропонують здійснювати оцінювання опінії відповідно до [12], що дасть змогу отримати ширший діапазон значень, в результаті чого результат буде більшим до достовірного.

У математичній моделі оцінювання опінії [12] враховано важливість коментарів за датою та автором. У такий спосіб після застосування лінгвістичних змінних розраховується важливість кожного коментаря відповідно до часу написання та автора.

Після врахування важливості кожного коментаря відбувається узагальнення лінгвістичних змінних та їхнє нормування відповідно до [12].

Наприкінці цього етапу автори отримують нормовані лінгвістичні змінні для кожного із компонент об'єкта.

1.2. Інтегральний показник компоненти об'єкта та об'єкта загалом

Наступним етапом, відповідно до інформаційної технології, є обчислення інтегрального показника компоненти об'єкта та об'єкта загалом. Детальний розрахунок описано авторами у роботі [12].

Описаний авторами метод дає змогу розраховувати вагові коефіцієнти між певними компонентами об'єкта. У результаті ми знаємо, що такий компонент, як “якість зображення”, може мати більшу вагу, ніж “ємність батареї” у об'єкта “фотоапарат”.

Вагові коефіцієнти розраховують за двома основними критеріями:

- за кількістю коментарів про певний компонент;
- за оцінкою експерта у цій предметній області.

Ці два критерії мають однакову вагу, об'єднуються, у такий спосіб визначається вага певної компоненти.

Розрахувавши вагові коефіцієнти, ми можемо обчислити числові значення кожної з компонент об'єкта та порівнювати об'єкти покомпонентно. Крім того, є можливість порівнювати

об'єкти загалом за всіма висловлюваннями користувачів за допомогою обчислення інтегрального показника об'єкта.

Важливість використання та застосування запропонованої авторами інформаційної технології оцінювання опінії полягає у такому:

1. Застосування лінгвістичних змінних, що уможливлює отримання числового значення опінії висловлювання ширшого діапазону, що, своєю чергою, впливає на збільшення точності та достовірності результату.

2. Розрахунок вагових коефіцієнтів, за допомогою яких можна обчислити інтегральний показник об'єкта, враховуючи важливість різних компонент об'єкта.

3. Розрахунок інтегрального показника об'єкта, що дасть змогу здійснювати порівняння об'єктів відповідно до числового значення їхнього інтегрального показника.

1.3. Вимоги до системи

Система оцінювання опінії повинна бути реалізована у вигляді web-скриптів, що забезпечить її зручну інтеграцію у будь-яку web-орієнтовану систему.

Розроблений web-додаток повинен бути простим та зручним у використанні, виводити результати обчислень за вхідними даними в максимальному інформативному вигляді для подальшого використання в практичній діяльності.

Враховуючи популярність систем управління контентом, у створюваній системі оцінювання опінії повинна бути передбачена можливість легкої інтеграції в CMS або інтернет-магазин чи іншу web-орієнтовану систему. При цьому кількість об'єктів та компонентів можна змінити під час інтеграції відповідно до вимог користувача, і треба мати змогу отримувати з розрахунків самого користувацького веб-сервісу.

2. Структура системи оцінювання опінії

Система оцінювання опінії, базована на інформаційній технології оцінювання опінії, повинна складатися з двох основних блоків:

1. Функціональний блок.
2. Бази знань.

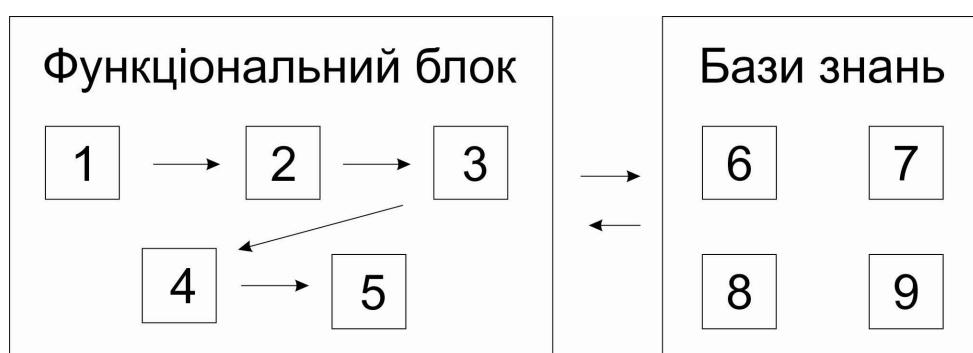


Рис. 2. Структурна схема системи оцінювання опінії

Функціональний блок – основний блок, котрий відповідає за опрацювання всіх вхідних даних та формування вихідної інформації. Функціональний блок повинен складатися з:

1. Блока збору інформації – призначений для збору інформації стосовно певного об'єкта оцінювання (виявлення об'єкта та його компонент). Для реалізації цього блока потрібне використання інформаційно-пошукових систем. Збір інформації блоком 1 здійснюється із використанням знань з блока 6.

2. Блока попередньої обробки – призначений для попередньої обробки текстової інформації, такої як поділ на частини мови, семантичний та синтаксичний аналіз речення. Для реалізації цього блока потрібно задіювати системи графоморфологічного аналізу тексту, описані в огляді існуючих рішень.

3. Блока оцінювання опінії – використовується для отримання числового значення позитивних та негативних висловлювань із застосуванням лінгвістичних змінних. Для оцінювання опінії блоком 3 використовуються блоки позитивних та негативних слів-опіній із блоків 8, 9.

4. Блока підрахунку – призначений для обчислення:

- важливості висловлювань за датою та відповідно до власника опінії;
- нормалізованих лінгвістичних змінних;
- числових значень опінії компонентів об'єкта;
- вагових коефіцієнтів компонентів об'єкта;
- інтегрального показника об'єкта.

Для підрахунку важливості висловлювань за датою та відповідно до власника опінії отримують дані із блока 7.

5. Блока виведення – використовується для виведення кінцевої інформації користувачу у зручній для нього формі. Для зручнішого сприйняття інформації використовуються як числові значення, так і діаграми.

Бази знань повинні зберігати таку інформацію:

6. База знань об'єктів та компонентів – містить інформацію про об'єкти дослідження та їхні компоненти.

7. База знань власників опінії – містить інформацію про важливість певних власників опінії щодо певних об'єктів. Також у цій базі знань повинна бути можливість надання важливості за датою.

8. База знань позитивних слів – містить позитивні слова-опінії.

9. База знань негативних слів – містить негативні слова-опінії.

Запропонована структура системи повинна використовувати інформаційну технологію оцінювання опінії. Авторами реалізовано 3-й, 4-й та 5-й блоки системи оцінювання опінії. Математичний формалізм реалізації цих блоків описано у [12].

3. Технології реалізації системи

Для реалізації 3-го, 4-го та 5-го функціональних блоків системи оцінювання опінії автори використали технології web-програмування:

1. Мову програмування PHP [14] – для реалізації всіх програмних елементів.

2. Google Charts API [15] – для реалізації виведення діаграм користувачу.

Мову програмування PHP використано з метою спрощення та надійності розробки, оскільки сьогодні вона підтримується переважною більшістю хостинг-провайдерів та є однією з найпопулярніших скриптових мов.

Блок виводу, крім того, що написаний на PHP, використовує Google Charts API. Chart API дає змогу створювати діаграмами з вхідних даних і вставляти їх у веб-сторінки. Нині підтримуються такі види графіків: лінійні, стовпчикові, площинні, кругові, графіки-карти, діаграми Венна та інші (загалом більше за 60).

Діаграми, побудовані за допомогою Google Chart API, реалізуються як класи JavaScript, що дає змогу легко налаштовувати кожен графік відповідно до вигляду web-інтерфейсу, на якому він буде застосовуватися. Крім того, ці графіки інтерактивні, що дозволяє використовувати їх для створення складних панелей та подій, інтегрованих з web-сторінкою. Відображення графіків відбувається за допомогою технології HTML5/SVG, що забезпечує сумісність з різними браузерами та платформами.

4. Функціонал системи

4.1. Вхідні та вихідні дані

Вхідними даними для блока оцінювання опінії є функції належності лінгвістичних змінних (відповідно до опитування користувачів чи методом поділу відповідно до шкали). На виході отримуємо числові значення опінії висловлювання відповідно до функцій належності.

Вхідними даними для блока підрахунку є кількість позитивних, негативних та нейтральних відгуків, а також числові значення опінії висловлювань. На виході отримуємо результат, котрий потрібно подати у доступній користувачеві формі. Для цього використовується блок виведення.

Вхідними даними для блока виведення є:

– коефіцієнти важливості компонент об'єкта (відповідно до кількості висловлювань та оцінки експерта);

– числові значення компонент об'єкта;

– інтегральний показник об'єкта;

На виході блока виведення отримуємо:

– діаграмами порівняння об'єктів покомпонентно;

– діаграмами порівняння інтегральних показників об'єктів.

Інтерфейс системи для вводу вхідних даних зображенено на рис. 3 [15].

Object 1

	Positive	Negative	Neutral
Feature 1	251	631	21
Feature 2	321	346	26
Feature 3	225	701	15
Feature 4	425	753	19

Object 2

	Positive	Negative	Neutral
Feature 1	612	1261	61
Feature 2	166	862	21
Feature 3	419	1212	59
Feature 4	519	1256	91

Рис. 3. Інтерфейс введення вхідних даних

4.2. Результат роботи

Після введення вхідних даних виводяться результати обчислень у числовому вигляді та у вигляді графіків Google Chart. Числові результати використання системи наведено на рис. 4.

Feature weight:

Weight of feature 1 = 27.62

Weight of feature 2 = 16.96

Weight of feature 3 = 25.61

Weight of feature 4 = 29.82

Рис. 4. Діаграма підрахунку вагових коефіцієнтів компонент об'єкта

На рис. 5 наведено результат порівняння об'єктів покомпонентно. У цьому прикладі зображене порівняння чотирьох компонентів двох об'єктів. Як видно з рисунка, Object1 має негативні коментарі за двома із чотирьох компонент, проте невелику кількість, Object2 – по одній із чотирьох, проте значну кількість. Водночас Object1 має позитивні коментарі щодо двох із чотирьох компонент, причому значно більше, ніж Object2, аналогічно і Object2 має позитивні коментарі за двома із чотирьох компонент, причому значно більше, ніж Object1. Постає питання: а який об'єкт кращий (релевантніший) з погляду кількості позитивних та негативних коментарів із урахуванням вагових коефіцієнтів компонент? Відповідь на це питання дає наступний рисунок.

Результати використання системи в графічному вигляді для порівняння інтегрального показника об'єктів наведено на рис. 6. Під час підрахунку інтегрального показника враховуються

вагові коефіцієнти компонент об'єкта. Отже, можна сказати, що Object1 релевантніший, ніж Object2, за кількістю оціночних відгуків про компоненти об'єктів із урахуванням вагових коефіцієнтів компонент.

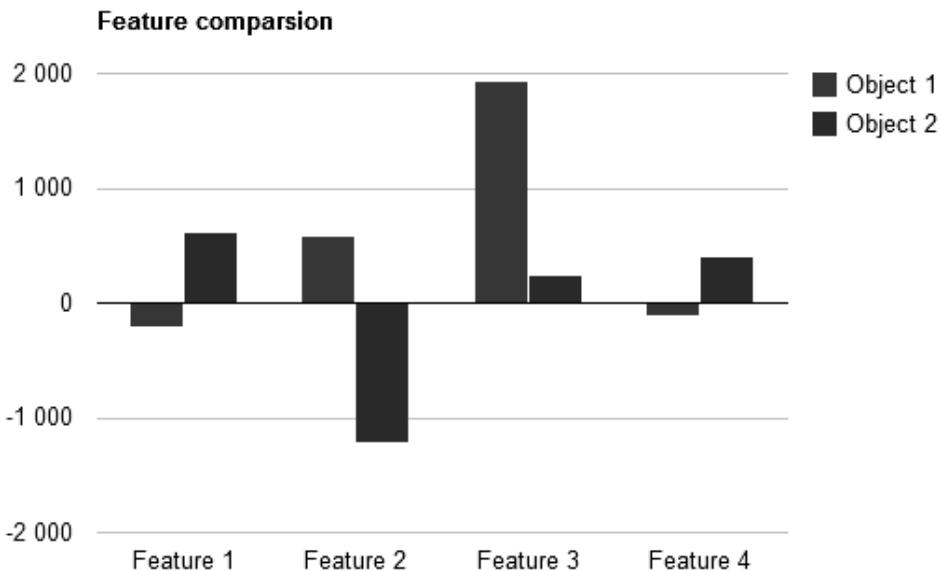


Рис. 5. Діаграма порівняння компонентів об'єкта

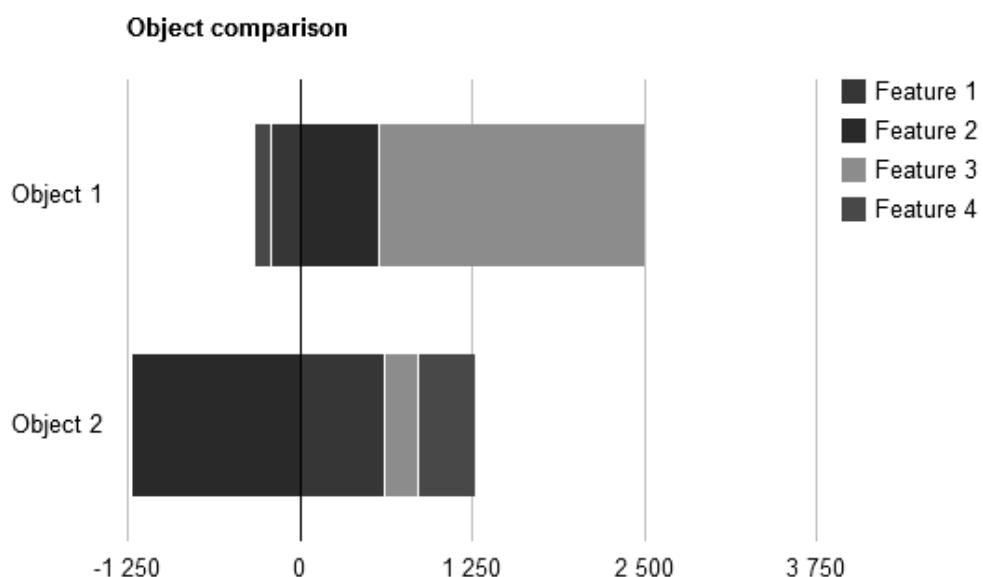


Рис. 6. Діаграма порівняння об'єктів

Висновки

У статті розроблено та описано інформаційну технологію та систему оцінювання опінії, завдяки якій можна проаналізувати опінії користувачів стосовно певного об'єкта.

Виконано огляд наявних рішень методів машинного навчання їх порівняння, що дало змогу визначити важливість семантичного підходу до аналізу неструктурованих текстів.

Розроблена інформаційна технологія оцінювання опінії, що дає змогу врахувати такі важливі етапи обчислення опінії, як семантичний аналіз речень та застосування лінгвістичних змінних із метою розширення діапазону можливих значень опінії об'єкта для покращення розрахунку узагальненої оцінки об'єкта.

Розроблена система оцінювання опінії на базі описаної інформаційної технології дає змогу розраховувати вагові коефіцієнти компонент об'єкта та інтегральний показник з метою порівняння декількох оцінюваних об'єктів.

На базі запропонованої інформаційної технології реалізовано та описано блоки оцінювання опінії та розрахунку інтегрального показника об'єкта, що дає можливість у візуальному інтерфейсі бачити показники об'єктів для їх подальшого порівняння.

1. Хомів Б. А. *Компаративний аналіз математичних моделей, методів та засобів оцінювання опінії в текстових даних інтернет-ресурсів* / Б. А. Хомів, С. А. Лупенко, А. С. Сверстюк // Вісник Хмельницького національного університету. – 2011. – № 6. – С. 7–16. 2. Котельников Е. В. Автоматический анализ тональности текстов на основе методов машинного обучения / Е. В. Котельников, М. В. Клековкина // Конференция “Диалог 2012”. – 30 мая – 3 июня 2012 г. [Електронний ресурс]. – Режим доступу : <http://www.dialog-21.ru/digests/dialog2012/materials/pdf/105.pdf>. 3. Хомів Б. А. Программа система оцінювання опінії текстової інформації із застосуванням байесівського класифікатора / Б.А. Хомів, С.А. Лупенко, О.А. Пастух / Матеріали міжнародної конференції “Системний аналіз та інформаційні технології”, Київ, 23–28 травня 2011 р. – К.: Інститут прикладного системного аналізу Національного технічного університету України “Київський політехнічний інститут”, 2011 – С. 224–225. 4. Сибиряков А. Извлечение мнений о товарах из форумов и блогов с учетом тональности / А. Сибиряков // II Российской летняя школа по информационному поиску RuSSIR'2008, 1–5 сентября 2008 г., Труды Второй Российской конференции молодых ученых по информационному поиску / [отв. за вып. А. Н. Каркищенко]. – Таганрог : Изд-во ТТИ ЮФУ, 2008. – С. 63–78. 5. Величко В. Автоматизированное создание тезауруса терминов предметной области для локальных поисковых систем / Величко В., Волошин П., Світла С. // “Knowledge – Dialogue – Solution” International Book Series “INFORMATION SCIENCE & COMPUTING”, Number 15. – FOI ITHEA Sofia, Bulgaria. – 2009. – Р. 24–31. 6. NLProcessor – Text Analysis Toolkit [Електронний ресурс]. – Режим доступу : <http://www.infogistics.com/textanalysis.html>. 7. Opinion Lexicon [Електронний ресурс]. – Режим доступу: <http://www.cs.iic.edu/~liub/FBS/opinion-lexicon-English.rar> 8. Синтаксический анализатор от Dictum [Електронний ресурс]. – Режим доступу : <http://www.dictum.ru/ru/syntax/blog> 9. Text Analysis Software Engine [Електронний ресурс]. – Режим доступу : <http://www.lexalytics.com/> 10. Синтаксический и семантический анализатор текстов ABBYY Compreno [Електронний ресурс]. – Режим доступу : <http://www.abbyy.ru/science/technologies/business/compreno/> 11. [Електронний ресурс]. – Режим доступу : SemanticForce 12. Хомів Б. А. Застосування лінгвістичних змінних та вагових коефіцієнтів при формуванні інтегральної оцінки об'єкта у задачах opinion mining / Б. А. Хомів, С. А. Лупенко, О. А. Пастух, Ю. В. Нікольський // Вісник Національного університету “Львівська політехніка”, “Комп’ютерні науки та інформаційні технології”. – 2012. – № 732. – С. 264–273 [Електронний ресурс]. – Режим доступу : http://ena.lp.edu.ua:8080/bitstream/ntb/14921/1/39_Lupenko_264_273_732.pdf . 13. PHP Programming Language [Електронний ресурс]. – Режим доступу : <http://php.com> 14. Google Charts [Електронний ресурс]. – Режим доступу : <https://developers.google.com/chart/> 15. Інформаційна система підрахунку інтегрального показника об'єкта та обчислення вагових коефіцієнтів [Електронний ресурс]. – Режим доступу : http://91.201.156.253/opinion_mining/mark/

Подяки

Автори висловлюють подяку Тернопільському національному технічному університетові ім. І. Пулюя за надану технічну можливість розташування сервера для реалізації системи оцінювання опінії; компанії SemanticForce (<http://semanticforce.net>) в особі директора Всеволода Гаврилюка за сприяння у розумінні потреб бізнес-сектору в цій предметній області та тестування функціонала реалізованих модулів системи оцінювання опінії.