ПРИКЛАДНІ ДОМЕНИ І ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ

УДК: 004.421

А.Б.Бєгунов

АЛГОРИТМ АВТОМАТИЗОВАНОГО ВИЗНАЧЕННЯ ЕМОЦІЙНОГО ЗАБАРВЛЕННЯ ТЕКСТОВИХ ДАНИХ

Національний технічний університет України «Київський політехнічний інститут»

кафедра. програмного забезпечення комп'ютерних систем

Науковий керівник – Заболотня Т.М., к.т.н.

*Дана стаття присвячена рішенню питання розробки алгоритму автоматизованого визначення емоційного забарвлення природномовних текстів. При цьому використано комбінований підхід з використанням класифікатора текстових даних і аналізу синтаксичних зв'язків між словами в тексті. Це дозволяє враховувати не тільки емоційне забарвлення (тональність) окремих слів, але і тональність груп синтаксично пов'язаних слів. Такі групи складаються зі слів, що мають емоційне забарвлення і безпосередньо беруть участь у формуванні тональності тексту, та слів, що підсилюють або послаблюють тональність інших слів в межах групи. Такий підхід передбачає ранжирування текстових даних за категоріями емоцій, а також розширюваність за рахунок можливості додавання нових к**ategорій емоційного забарвлення текстових даних.***
***На основі даного подходу ррозроблено алгоритм автоматизованого** визначення емоційного забарвлення природномовних текстових даних.*

Данная статья посвящена решению вопроса разработки алгоритма автоматизированного определения эмоциональной окраски текстов на естественном языке. Для решения поставленной задачи предложено использовать комбинированный подход с использованием классификатора текстовых данных и анализа синтаксических связей между словами в тексте. Это позволяет учесть не только эмоциональную окраску (тональность) отдельных слов, но и тональность групп синтаксически связанных слов. Такие группы состоят из слов, имеющих эмоциональную окраску и непосредственно принимающих участие в формировании тональности текста, и слов, усиливающих или ослабляющих тональность других слов в пределах группы.
Такой подход предусматривает ранжирование текстовых данных по категориям эмоций, а также расширяемость за счет возможности добавления новых категорий эмоциональной окраски текстовых данных.
На основе данного подхода разработан алгоритм автоматизированного определения эмоциональной окраски текстов на естественном языке.

This article is devoted to solving the problem of the automated computation of emotional coloring of natural language text. In this case a combined approach is used. This approach provides the categorization of text data and the analysis of syntactic relations between words in this text. It enables to consider the emotional color (tonality) of individual words, but the key groups of syntactically related words. These groups are composed of emotionally colored words and words that amplify or weaken the emotional coloring of other words within the group.
This approach provides the ranking of text data into categories of emotions and enables to add new categories of emotions. An algorithm of the automated computation of emotional coloring of natural language text based on this approach is developed.

Ключові слова: природномовні текстові дані, емоційне забарвлення текстових даних, тональність текстових даних, автоматизована обробка текстових даних, алгоритм класифікації, синтаксичний аналіз

Вступ

Однією з важливих переваг використання інформаційних технологій є можливість автоматизованого аналізу великих масивів даних. Спектр завдань, які відносяться до інформатизації різних сфер життя, є надзвичайно широким і включає, зокрема, отримання інформації з метою прийняття рішень, навчання, розв'язання наукових та організаційних задач тощо [1].

На сьогоднішній день із зростанням кількості сфер застосування інформаційних технологій збільшується необхідність у різного роду програмних засобах автоматизованої обробки даних різного типу (пошук, вилучення інформації, структурування тощо). Великий відсоток даних, що використовуються, припадає на природномовні текстові дані.

Важливим напрямком досліджень у галузі автоматизованого аналізу природномовних тек-

стових даних є оцінювання тональності, або емоційного змісту, заданого тексту. Для вирішення багатьох задач, пов'язаних з обробкою текстових даних, важливим є не тільки врахування формальних характеристик тексту, а і його емоційного забарвлення і, як наслідок, психологічного впливу на людину. Під тональністю тексту зазвичай розуміють позитивне, негативне або нейтральне забарвлення як цілого текстового документу, так і його окремих частин, які мають відношення до певних понять, таких як персони, організації, бренди тощо [1]. Існуюче програмне забезпечення реалізує автоматизований аналіз емоційного змісту тексту виключно за ступенем позитивності, що робить результати роботи таких програм досить грубими і значною мірою звужує коло сфер їх застосування. Наприклад, якщо текстове повідомлення негативне, то воно може виражати страх або агресію, що не є тотожним емоційним забарвленням з точки зору людини. Виправити дану ситуацію можна шляхом розширення спектра тональностей, які здатна розрізняти програмна система. Такий підхід може набути широкого застосування у різних сферах: соціальні мережі, психологія, маркетинг тощо.

Отже, розробка програмного забезпечення для визначення багатовекторної картини емоційного забарвлення тексту є цікавою та актуальною задачею.

Основні визначення

Перш, ніж сформулювати мету та задачі даного дослідження, дамо визначення основних понять, якими будемо оперувати у статті.

Природномовними текстовими даними (текстом) будемо називати сукупність речень будь-якою природною мовою [1].

Під *емоційним змістом (тональністю)* будемо розуміти певну емоційну забарвленість тексту, яка формується тональністю його емоційно забарвлених складових одиниць та правил їх поєднання [2], що визначає належність тексту до певної категорії емоцій, наприклад:

- радість;
- страх;
- задоволеність;
- агресія.

Кожна така категорія має певні характеристики, за якими вона може бути ідентифікована. Ці характеристики відображаються у певних параметрах тексту, що належить до цієї категорії. Як наслідок, довільний текст з певними параметрами може бути віднесений до відповідної категорії. Емоційну забарвленість тексту визначають такі параметри:

- емоційно забарвлені слова (терми), що належать до певної емоційної категорії;
- відношення (зв'язки) між цими словами (термами) та (або) іншими словами в тексті, що виникають в будь-якому природномовному тексті. Такі зв'язки носять синтаксично-семантичний характер і є невід'ємним елементом будь-якого речення природною мовою. Наявність таких відношень відображається на емоційній забарвленості всього тексту в цілому.

Огляд існуючих програмних рішень щодо визначення тональності текстів

1. Інтернет-ресурс «TextTools.RU» має інструмент, що проводить аналіз тексту і в якості результату визначає відсоток позитивного і негативного забарвлення тексту. Точність аналізу залежить від обсягу тексту [3].

Серед переваг програмного засобу «TextTools.RU» можна зазначити можливість визначення відсотку емоційного забарвлення, а не чітке віднесення до певної емоційної категорії. Недоліками є наявність лише двох категорій емоційного забарвлення (позитивне і негативне), неможливість розширення системи користувачем за рахунок додавання нових категорій та залежність якості аналізу від обсягу тексту.

2. Програмна система ВААЛ розроблена з метою автоматизованого прогнозувати ефекту неусвідомлюваного впливу текстів на масову аудиторію [4]. Для кожного виду психоемоційного впливу в системі передбачена своя категорія. Ступінь належності тексту до певної категорії подано шкалою оцінки даного виду впливу. Кожна шкала задається двома протилежними за змістом поняттями (наприклад, хороший – поганий, гарний – відштовхуючий, легкий – важкий і т.п.). Такий функціонал не повністю відповідає вирішенню задачі визначення емоційного забарвлення природномовних текстових даних , але є достатньо близьким до неї. В якості алгоритмічної основи у системі ВААЛ використовуються алгоритми фоносемантичної оцінки текстових даних. В основі оцінювання емоційного впливу на підсвідомість людини тексту російською мовою лежить алгоритм Журавльова. Для україномовних текстів використовується алгоритм Левицького. Ці алгоритми базуються на аналізі емоційного впливу звуків в окремих словах. Тобто основною одиницею, що формує вплив тексту на підсвідомість людини в даному випадку є звук. Для розв'язання задачі автоматизованого визначення емоційного забарвлення такий підхід не є коректним. Емоційне забарвлення тексту формується за допомогою емоційно забарвлених слів, а також зв'язків між ними та нейт-

рально забарвленими словами. Тобто, елементарною одиницею, що робить внесок у ту чи іншу емоційну забарвленість, повинно бути слово.

Основними перевагами програмної системи ВААЛ є:

- ранжирування не тільки за категоріями психоемоційного впливу, але й за ступенем впливу в межах кожної з категорій;
- великий набір категорій емоційного впливу;

Недоліками системи є:

- неможливість розширення системи користувачем за рахунок додавання нових категорій.

Постановка задачі

В контексті описаної проблематики та на основі результатів аналізу розглянутих вище програмних рішень можна сформулювати такі вимоги до функціональних можливостей програмного забезпечення визначення емоційного забарвлення тексту:

1. Розширюваність.

Суттєвим недоліком розглянутих рішень є відсутність можливості додавання користувачем категорій, за якими проводиться категоризація. Це робить вимогу розширюваності доцільною та обґрунтованою.

2. Ранжирування.

Будь-які природномовні тексти можуть мати глибокий складний емоційний зміст. Тому в загальному випадку будь-який текст може бути частково віднесений до кількох категорій одразу. Отже, задача зводиться до ранжирування текстових даних за декількома певними категоріями.

3. Глибокий аналіз тексту для врахування всіх параметрів категорій.

Для підвищення точності аналізу текстових даних потрібно врахувати всі фактори, які визначають результат категоризації. Такими факторами є характеристики (параметри) категорій, за якими відбувається категоризація. Як було зазначено вище, кожна категорія має два таких параметри: терми, що входять до кожної з категорій та зв'язки між термами.

Таким чином, виходячи із наведених вище аргументів, **метою** даного дослідження стало розширення функціональних можливостей програмного забезпечення автоматизованого визначення емоційного забарвлення природномовних текстових даних та підвищення точності аналізу цих даних за рахунок розробки нового алгоритму визначення тональності природномовних текстів

Опис розробленого алгоритму

Поставлена задача є задачею нечіткої класифікації (ранжирування): потрібно знайти ступінь належності вхідних даних до кожної із заздалегідь заданих категорій.

Нехай емоційний зміст тексту визначено набором категорій (класів) емоцій (наприклад, страх, задоволення тощо): $C_1, C_2, \ldots, C_m$. Текст подається набором документів $D = \{d_1, d_2, \ldots, d_n\}$. Для кожної категорії будується функція статусу класифікації $CSV_j(d_i) \in [a;b]$, де a та b – мінімальне та максимальне значення ступеня входження i-го документу до категорії C_j. Потрібно побудувати алгоритм F, що забезпечить ранжирування [1]:

$$F(d_i, C_1, C_2, \ldots, C_m) = \{CSV_1(d_i), CSV_2(d_i), \ldots, CSV_m(d_s)\} \quad (1)$$

Класичний алгоритм розв'язання такої задачі складається з таких кроків:

1. Створення навчальної колекції документів – масиву текстів, для яких вже відомі значення ступеня входження до кожної з категорій.

2. Навчання класифікатору на навчальній колекції. Вилучення з навчальних колекцій набору слів, які є ключовими для процесу класифікації. Визначення ваги $w_i \in [a;b]$ – ступеня входження k-го слову до даної категорії.

3. Класифікація (ранжирування) вхідного тексту:

(a) розбиття тексту на набір термів;

(b) співставлення кожного терму та даних, що були сформовані під час навчання;

(c) врахування внеску ваги терму в значення *CSV* для кожної категорії.

4. Повернення результату.

У нашому випадку термами слугуватимуть слова. В залежності від їх емоційної забарвленості вони будуть робити внесок у значення ступеня належності до певної категорії. Відрізок $[a;b]$ в доцільно обрати як [0;1], що є, зазвичай, дуже зручним з точки зору подання даних.

Описаний алгоритм має суттєвий недолік. Виходячи з означення емоційних категорій, очевидно, що алгоритм аналізу ступеня належності текстових даних до тієї чи іншої категорії слід будувати, спираючись на параметри, якими характеризуються ці категорії: певні емо-

ційно забарвлені слова та зв'язки між словами в тексті. Але описані вище кроки враховують лише самі слова і не враховують зв'язків між ними. Отже, для розв'язання поставленої задачі потрібен більш глибокий аналіз вхідного тексту.

Виходячи із зазначеного вище, доцільно проводити класифікацію текстових даних за категоріями в два етапи: аналіз зв'язків між словами і врахування наявності емоційно забарвлених слів.

1. **Аналіз зв'язків між словами.**

Розглянемо, яким чином зв'язки можуть вплинути на емоційний зміст тексту [5].

Нехай для певної категорії C_k терм t_i має вагу w_i.

Виділимо множини термів $A_1=\{\alpha^1_1,\alpha^1_2,\dots\alpha^1_{p0}\}$ і $A_0=\{\alpha^0_1,\alpha^0_2,\dots\alpha^0_{p1}\}$, що синтаксично пов'язані з t_i. Зазначимо, що важливим є не тип або характер зв'язків, а лише характер самої взаємодії. Терми з множин A_1 і A_0 не мають безпосереднього впливу на емоційне забарвлення тексту відносно C_k, але будь-який терм з A_1 посилює емоційне забарвлення t_i, а будь-який терм з A_0 його послаблює.

Тепер візьмемо множини термів $A_{10}=\{\alpha^{10}_1,\alpha^{10}_2,\dots\alpha^{10}_{p10}\}$ та $A_{11}=\{\alpha^{11}_1,\alpha^{11}_2,\dots\alpha^{11}_{p11}\}$, що, відповідно, послабляють та посилюють вплив множини A_1. Аналогічно введемо множини термів A_{00} та A_{01}, що послабляють та посилюють вплив термів з множини A_0. Назвемо всі такі множини та терми, що до них входять, множинами та термами впливу.

Таким чином, можна сформувати багаторівневу структуру множин таких термів, що, відповідно, посилюють або послаблюють вплив термів з множин попереднього рівня. Послаблення чи посилення емоційного забарвлення терму t_i виражається в модифікації його ваги w_i з використанням коефіцієнтів, що ставляться у відповідність термам з множин A_1, A_0, A_{11}, A_{10}, A_{01},A_{00},…. Кожній такій множині поставимо у відповідність множину коефіцієнтів Q_1, Q_0, Q_{11}, Q_{10}, Q_{01},Q_{00},…. Отже, при врахуванні впливу всіх можливих термів на t_i матимемо уточнену вагу для t_i:

$$w_i'=\gamma\,(Q_1, Q_0, Q_{11}, Q_{10}, Q_{01}, Q_{00},\dots,w_i). \qquad (2)$$

Коефіцієнти для термів впливу можуть бути визначені на етапі навчання з використанням спеціальних документів, для яких заздалегідь відомо, який вплив мають ті чи інші терми.

2. **Аналіз емоційно забарвлених слів.**

На цьому етапі обчислюються значення *CSV* по кожній з категорій для чергового документу. З попереднього етапу для кожного терму в документі маємо його скореговану вагу. Далі, в залежності від обраного методу класифікації, вага кожного з термів враховується при обчисленні ступеня належності документу до кожної з категорій. В якості класифікатору в розробленому програмному засобі обрано метод наївної байесовської класифікації [6]. Для наївної байесовської моделі робиться припущення про статистичну незалежність термів (характеристик документів) t_1, $t_2,\dots,$ t_n. На основі документів з навчальної колекції визначаються коефіцієнти, що характеризують умовну ймовірність належності слова t_i до відповідної категорії. Тобто, вага терму (слова) w_i' має імовірнісний зміст. Тоді значення *CSV* розраховується за формулою:

$$\boldsymbol{CSV}_j(d_k)=\frac{\prod_{i=1}^{n} w_i'}{\prod_{i=1}^{n} w_i' + m\cdot\prod_{i=1}^{n}(1-w_i')}, \text{ де } m - \text{кількість категорій.} \qquad (3)$$

Таким чином, алгоритм автоматизованого визначення емоційного забарвлення текстових даних, що пропонується у даній статті, має такий вигляд:

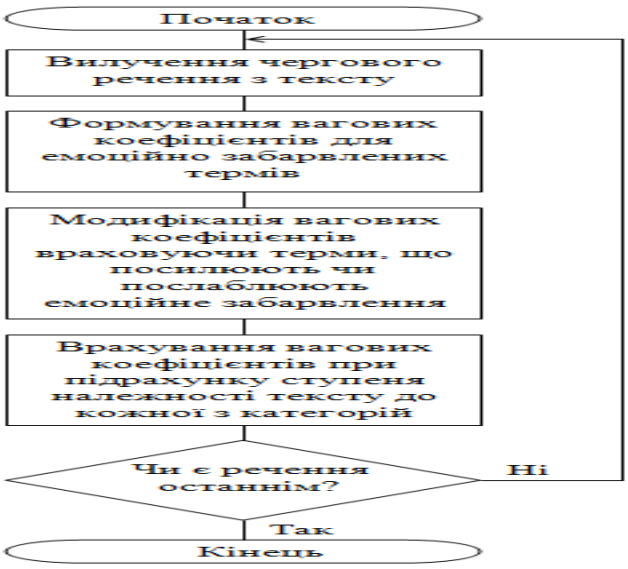

Рис.1. Схема алгоритму визначення емоційного забарвлення природномовних текстових даних

Висновки

На основі класичного підходу до ранжирування текстових даних розроблено модифікований алгоритм нечіткої класифікації природномовних текстових даних, який дозволяє більш точно визначати емоційне забарвлення останніх, ніж існуючі алгоритми, а також робить механізм для управління категоріями емоцій (їх додаванням та видаленням) більш гнучким.

Подальше вивчення питання видається автору перспективним, оскільки розроблений алгоритм може бути застосований як основа системи автоматизованого визначення емоційного змісту природномовних текстових даних, яка може бути використана при вирішенні широкого спектру задач, зокрема, для комп'ютеризованого аналізу впливу інформації із ЗМІ на людей, аналізу психоемоційного стану колективу у крупних корпораціях тощо.

Використана література

1. Ландэ, Д.В. Интернетика. Навигация в сложных сетях: модели и алгоритмы [Текст] / Д.В. Ландэ, А.А. Снарский, И.В. Безуднов. — М.: Либроком, 2009. — 264с.

2. Гаспаров, Б. М. Язык, память, образ. Лингвистика языкового существования [Текст] / Б. М. Гаспаров. — М.: Новое Литературное Обозрение,1996. — 352 с.

3. TextTools [Електронний ресурс]. — Режим доступу : http://texttools.ru/

4. ВААЛ [Електронний ресурс]. — Режим доступу : http://vaal.ru/

5. Леонтьева, Н. Н. Автоматическое понимание текста: системы, модели, ресурсы [Текст] / Леонтьева, Н. Н. — М.: Издательский центр «Академия», 2006. — 304 с.

6. David D. Lewis. Naive (Bayes) at forty: the independence assumption in information retrieval. [Текст] / D. Lewis David. In Proceedings of the ECML-98, Chemnitz, DE: Springer Verlag, Heidelberg, DE., 1998. —40 с.

Відомості про авторів

Бєгунов Андрій Борисович - магістрант, кафедри програмного забезпечення комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут», коло наукових інтересів: розвиток теорії алгоритмів та обчислень.
E-mail: arxton@mail.ru

Заболотня Тетяна Миколаївна, к.т.н., старший викладач кафедри програмного забезпечення комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут», коло наукових інтересів: об'єктно-орієнтоване програмування, розвиток теорії алгоритмів та обчислень, розроблення методів програмування для динамічних середовищ та агентно-орієнтованого програмування
E-mail: tatiana104@yandex.ru