УДК 621.396.967

к.т.н., доц. **Пампуха І.В.** (ВІКНУ)
к.військ.н. **Нікіфоров М.М.** (ВІКНУ)
к.т.н. **Лоза В.М.** (ВІКНУ)
Ступницька О.І. (ВІКНУ)

ОБҐРУНТУВАННЯ ВИБОРУ РАЦІОНАЛЬНОГО АЛГОРИТМУ АНАЛІЗУ ТОНАЛЬНОСТІ РІЗНОМОВНОЇ ТЕКСТОВОЇ ІНФОРМАЦІЇ ДЛЯ ЗАДАЧІ МОНІТОРИНГУ ІНФОРМАЦІЙНОГО ПРОСТОРУ

У роботі проведено аналіз алгоритмів та методик, що здійснює автоматизацію процесу визначення тональності текстової інформації. Проведено обґрунтування вибору раціонального алгоритму аналізу тональності різномовної текстової інформації для задачі моніторингу інформаційного простору з метою виявлення джерел інформаційного впливу і запропоновано комбінований метод, який поєднує переваги методу SVM та методу ключових слів без отримання додаткових недоліків. Визначено що комбінований метод автоматичного визначення тональності тексту, що об'єднує результати двох потужних класифікаторів (SVM та на основі ключових слів), дає можливість досягти підвищення якості класифікації в порівнянні не лише з цими класифікаторами, а й з найкращими результатами інших підходів. Виникнення такого ефекту забезпечується за рахунок врахування степені впевненості класифікаторів у своїх рішеннях та підбору оптимальних параметрів на основі ковзного контролю.

Ключові слова: тональність, алгоритм, текстова інформація, моніторинг, інформаційний простір.

Вступ. Одним із основних методів ведення інформаційних операцій є інформаційний вплив, що надається з метою інформаційного керування – непрямого інформаційного впливу,

коли об’єкту керування дається певна інформаційна картина, під впливом якої він формує лінію своєї поведінки.

У XXI столітті вплив на інформаційно-комунікаційні технології - ресурси супротивника в ході інформаційної операції може стати найпродуктивнішим знаряддям інформаційних війн.

Важливою характеристикою інформації як сукупності пов’язаних змістом слів – є тональність. Аналізуючи тональність текстів, можна встановити багато характеристик інформації.

На сьогоднішній день аналіз тональності текстів є перспективним напрямком досліджень для багатьох організацій, науковців, компаній і держави в цілому. При якісному аналізі емоційних елементів можна проводити точніший тональний аналіз текстових повідомлень. Сьогодні внаслідок наповнення соціальних мереж Інтернету величезною кількістю текстового матеріалу задача аналізу емоційного забарвлення тексту є також інструментом соціологічного дослідження і знаходить величезне поле для свого застосування.

Аналіз останніх досліджень і публікацій, постановка задачі. У забезпеченні результату політичного аналізу й прогнозування вирішальну роль відіграє оптимізація методів, які становлять арсенал наукового напряму. Широкого визнання набули методи моделювання. Різноманітні моделі використовували такі відомі теоретики, як С. Прайс, Д. Бартолом'ю (модель соціальної мобільності); Ван Гіг, П. Чекленд (модель м'яких систем); Дж. Річардсон (модель гонки озброєнь); Д. Істон, Г. Спіро, Г. Алмонд, Т. Парсонс (моделі функціонування політичних систем). [1-3]

Інформаційне протиборство, на думку американських фахівців корпорації “Ренд”, являє собою використання державами глобального інформаційного простору й інфраструктури для проведення стратегічних воєнних операцій і зменшення впливу на власний інформаційний ресурс[5-6]. Однією з характерних тенденцій, яка склалася в сучасних умовах, не тільки в Україні, а й у світі, – це випереджальний розвиток форм, способів, технологій і методик впливу на свідомість (підсвідомість), психологію й психічний стан людини в порівнянні з організацією протидії негативним, деструктивним психологічним впливам, інформаційно-психологічним захистом особистості й суспільства в цілому.

Мета роботи: Обґрунтувати вибір раціонального алгоритму аналізу тональності різномовної текстової інформації з метою вибору оптимальної задачі моніторингу інформаційного простору.

Викладення основного матеріалу. В умовах великого поширення інформації серед населення та зокрема у засобах масової інформації, необхідно правильно її аналізувати. Для того, щоб робити це без помилково, необхідно використовувати програмні засоби, як такі, які забезпечують максимальну точність навіть при великій кількості інформації яку потрібно проаналізувати.

Проведемо обґрунтування вибору раціонального алгоритму аналізу тональності різномовної текстової інформації для задачі моніторингу інформаційного простору, та обґрунтуємо вибір характеристик, які будемо вважати визначальними для вибору алгоритму:

1. Першою та найголовнішою характеристикою буде наявність фази навчання. Це необхідно, аби забезпечити можливість врахувати всі нові або непомічені раніше особливості задачі. Якщо це не передбачити, то з появою нових чинників, кожного разу треба буде змінювати наявний чи обирати інші алгоритми. Так наприклад, розвиток технологій, який може мати досить не передбачений (інноваційний) характер, може змінити вплив певних типів даних.

2. Іншим прикладом може стати зміна політичного становища, що звичайно, змінює ставлення людей до тих чи інших подій (нажаль, таких прикладів в нас більш ніж достатньо). Тож, другою важливою характеристикою буде те, що навчання повинно відбуватися під контролем людини, аби вловлювати зміну людського ставлення до тих, чи інших подій. До того ж, машинне навчання з вчителем є більш розповсюдженим для аналізу текстів, в той час як навчання без вчителя є більш природним для навчання в біологічних чи фізичних системах, де немає впливу людського чинника в аналізі результату.

3. Третьою характеристикою є те, що цей метод повинен мати теоретичне обґрунтування. Це потрібно, аби мати можливість використовувати результати цього алгоритму (хоча б з певною похибкою) для інших досліджень. Ця характеристика суттєво звужує вибір раціонального алгоритму для задачі моніторингу, відкинувши всілякі евристики.

4. Напевно, останньою важливою характеристикою буде швидкість роботи, бо таку іншу важливу характеристику, як точність, можна поліпшити за рахунок навчання.

З огляду на всі вищезазначені характеристики ми обрали комбінований метод, який полягає в застосуванні двох потужних методів:

1. метод опорних векторів (support vector machine, SVM),
2. метод ключових слів.

Вони належать до методів машинного навчання з вчителем, більш того вони мають достатнє теоретичне обґрунтування. Метод опорних векторів є одним з найшвидших для знаходження вирішальних функцій, а його недолік з чутливістю до шумів компенсується методом ключових слів [4].

Опишемо, яким чином ми комбінуємо ці два метода. Отримавши результат роботи методів, ми обираємо гіпотезу про тональність тексту за певною стратегією:

1) жоден з методів не визначив клас (позитивна чи негативна ознака) – відносили текст до найбільш позитивного класу в даній задачі;

2) клас, визначений тільки в одному з методів – відносили до даного класу;

3) обидва методи визначили класи – тут можливі наступні варіанти:

• є збіг відповідей одного з класифікаторів SVM з відповідями методу ключових слів – відносили відгук до цього класу;

• вага класу в методі ключових слів перевищувала заданий поріг (визначений емпірично) – відносили відгук до цього класу;

• жодна з попередніх умов не виконувалась – приписували найбільш позитивну оцінку SVM.

Наведемо основний алгоритм спочатку методу опорних векторів, а потім методу ключових слів за словником.

Використовуємо скінчений автомат, за допомогою якого виконуємо етап попередньої обробки тексту. Отримуємо на вхід текст, який треба аналізувати. На виході з автомату отримуємо набір таблиць: таблиці розділів, таблиці речень та лексем.

Нехай задано скінчений автомат $(Q, \Sigma, \delta, S_0, F)$, де Q - множина станів, Σ - алфавіт, δ - функція переходу, $S_0 \in Q$ - початковий стан, $F \subseteq Q$ множина кінцевих станів.

Опишемо побудову скінченого автомата. Назвемо множиною станів Q ті слова в тексті, які несуть інформацію з певною емоційною характеристикою (позитивна, негативна). Множина станів слів з емоційним забарвленням будується протягом навчання синтаксичного аналізатора.

Алфавітом мови Σ , яку розпізнає автомат - це кінцевий набір слів з множини всіх слів всіх текстів.

Значеннями функції переходу δ скінченого автомата є перехід від речення до речення, від лексеми до лексеми, від слова до слова. Від набору слів до лексем, та речень.

Початковий стан $S_0 \in Q$ є перше слово тексту.

Множина кінцевих станів $F \subseteq Q$ є набір таблиць з слів які мають певну емоційну забарвленість.

Для мінімізації скінченого автомата сформулюємо наступну теорему.

Для довільного скінченого автомата може буди побудовано еквівалентний йому скінчений автомат з найменшою кількістю станів.

Опишемо алгоритм, заснований на цій теоремі.

1. Побудова еквівалентного початковому детермінований автомат.

2. Якщо в отриманому скінченому автоматі залишились стани, недосяжні з початкової вершини, видаляємо їх.

3. До отриманого кінцевого автомату застосуємо алгоритм побудови розбиття множини станів на класи еквівалентності по відношенню $\equiv$ та побудуємо мінімальний кінцевий автомат M', як описано вище.

В ході роботи автомату, символи, що поступають на ввід, накопичуються в буфері. У визначених станах скінченого автомату, відбувається запис поточного складового буферу в одну з таблиць, після чого буфер оновлюється. Робота продовжується до досягнення кінцевого стану.

На етапі формування словника, відбувається визначення загальних ознак для цього тексту: виділення слів і словосполучень в тексті, що зустрічаються більше 3 разів, вважаються ознаками словника.

Параметри методу:

Маються на увазі ті ключові слова, які будуть використовуватись у методі ключових слів. Ці ключові слова розробники програмного коду задають під час створення програми. Дані параметри також можуть доповнювати або змінювати користувачі.

На етапі зважування для кожної ознаки зі словника вираховується і зберігається глобальна вага, що визначає значимість ознаки для розв'язку задачі аналізу тональності тексту. Для кожного слова в словнику рахуємо кількість текстів, в яких зустрічається це слово. Для кожного слова в словнику рахуємо кількість входжень цих слів даний текст. Для кожного слова формуємо векторну модель.

У векторній моделі текст представляється у вигляді вектора, кількість компонентів N якого збігається з кількістю ознак у словнику.

Кожен компонент є вагою відповідної ознаки в даному тексті. Для розрахунку ваги кожного елементу використовується підхід TF-IDF.

Етап навчання SVM-класифікатора - побудування в N-вимірному просторі ознак гіперплощин, кожна з яких розділяє вектори двох класів.

Проблема побудови оптимальної гіперплощини, що поділяє, зводиться до мінімізації $|W|$,(де вектор W – перпендикуляр до гіперплощини, що поділяє). Тоді задача квадратичної оптимізації, має вигляд:

$$\begin{cases} \|w\|^2 \to \min \\ c_i(w \cdot x_i - b) \geq 1, \quad 1 \leq i \leq n. \end{cases} \tag{1}$$

За теоремою Кунна-Такера ця задача еквівалентна двоїстій задачі пошуку точки функції Лагранжа

$$\begin{cases} \{L(w,b;\lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i (c_i((w \cdot x_i) - b) - 1) \to \min_{w,b} \max_{\lambda} \\ \lambda_i \geq 0, \quad 1 \leq i \leq n \end{cases} , \tag{2}$$

де $\lambda = (\lambda_1, ..., \lambda_n)$ – вектор змінних.

Зведемо задачу до еквівалентної задачі квадратичного програмування, яка включає в себе лише двоїсті змінні:

$$\begin{cases} -L(\lambda) = -\sum_{i=1}^{n} \lambda_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j c_i c_j (x_i \cdot x_j) \to \min_{\lambda} \\ \lambda_i \geq 0, \ 1 \leq i \leq n \\ \sum_{i=1}^{n} \lambda_i c_i = 0. \end{cases} \tag{3}$$

Нехай ми вирішили дану систему, тоді w та b можна знайти за формулами:

$$w = \sum_{i=1}^{n} \lambda_i c_i x_i \tag{4}$$

$$b = w \cdot x_i - c_i, \ \lambda_i > 0. \tag{5}$$

Отже, алгоритм класифікації може бути розписаний наступним чином:

$$a(x) = siqn\left(\sum_{i=1}^{n} \lambda_i c_i x_i \cdot x - b \right). \tag{6}$$

При цьому сумування йде не по всій вибірці, а лише по опорним векторам, де $\lambda_i \neq 0$.

На етапі налаштування параметрів відбувається пошук оптимальних параметрів для методу ключових слів та для алгоритму комбінації результатів класифікації. Цей пошук відбувається відповідно до параметрів, заданих програмістами та користувачем.

В результаті виконання двох методів отримаємо для кожного з них результати зі значенням Positive (текст або окремі його частини мають позитивну тональність) або Negative (текст або окремі частини тексту мають негативну тональність), залежно від складу тексту. Комбінуємо ці результати та отримуємо результат комбінованого методу.

Tf-IDF – статичний показник, що використовується для оцінки важливості слів у контексті документу, що є частиною колекції документів чи корпусу. Вага (значимість) слова пропорційна кількості вживань цього слова у документі, і обернено пропорційна частоті вживання слова у інших документах колекції.

TF (частота слова) – відношення числа входжень обраного слова до загальної кількості слів документу. Таким чином, оцінюється важливість слова t_i в межах обраного документу. Термін був введений Карен Спарк Джонс.

$$TF = \frac{n_i}{\sum_k n_k}, \tag{7}$$

де n_i є ччисло входжень слова в документ, а в знаменнику – загальна кількість слів в документі.

IDF (обернена частота документу) – інверсія частоти, з якою слово зустрічається в документах колекції. Використання IDF зменшує вагу широковживаних слів.

$$IDF = \log \frac{|D|}{|(d_i \supset t_i)|}, \tag{8}$$

де $|D|$ – кількість документів колекції;

$|(d_i \supset t_i)|$ – кількість документів, в яких зустрічається слово t_i (коли $n_i \neq 0$).

Вибір основи логарифму у формулі не має значення, адже зміна основи призведе до зміни ваги кожного слова на постійний множник, тобто вагове співвідношення залишиться незмінним. Іншими словами, показник TF-IDF це добуток двох множників: TF та IDF.

$$TF - IDF = TF \cdot IDF \ , \tag{9}$$

де, TF - частота слова,

IDF - обернена частота документу.

Більшу вагу TF-IDF отримають слова з високою частотою появи в межах документу та низькою частотою вживання в інших документах колекції.

Документ у векторній моделі розглядається як не впорядкована множина термів. Термінами в інформаційному пошуку називають слова, з яких складається текст, а також такі елементи тексту, як, наприклад, 2010, II-5 або Тянь-Шань. Різними способами можна визначити вагу термів в документі – «важливість» слова для ідентифікації цього тексту. Наприклад, можна просто підрахувати кількість вживань термів в документі, так звану частоту терми, – чим частіше слово зустрічається в документі, тим більша у нього буде вага. Якщо терм не зустрічається в документі, то його вага в цьому документі дорівнює нулю.

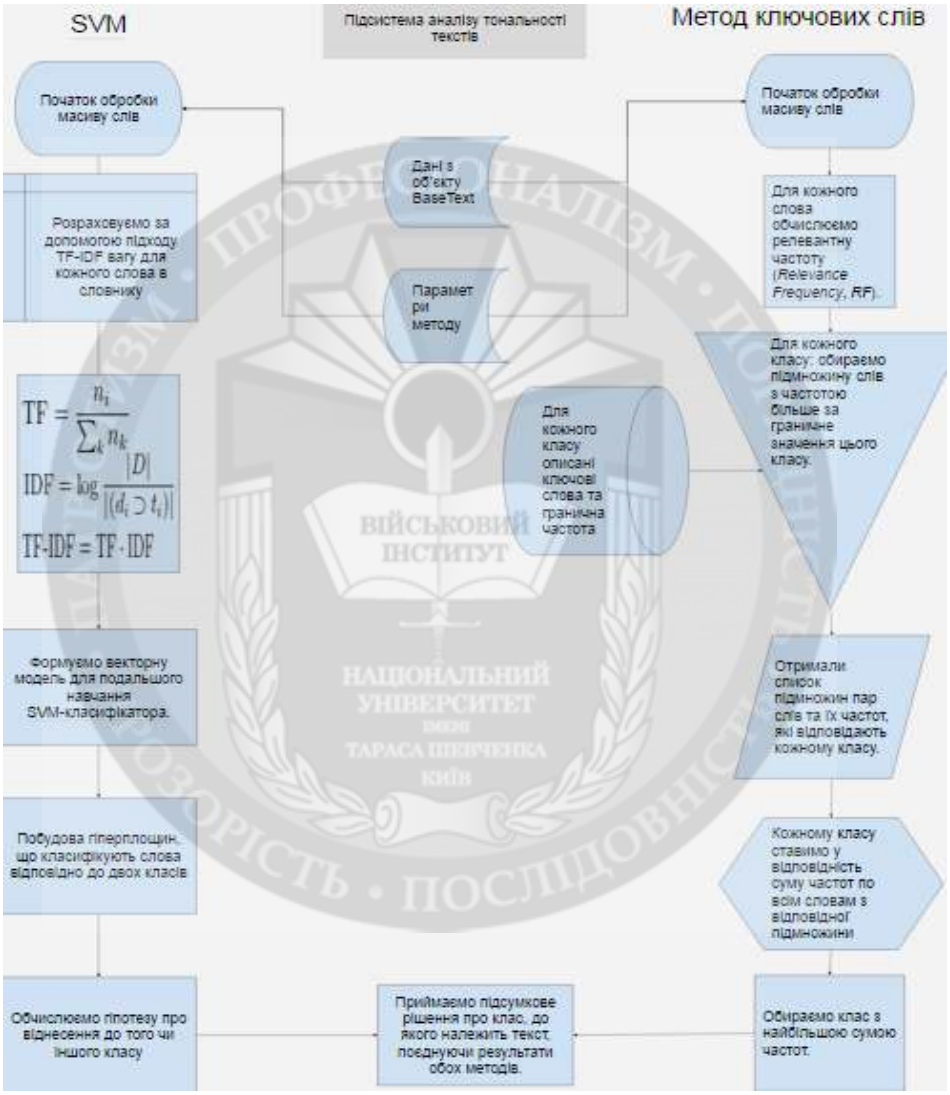

Рис. 1. Реалізація комбінованого методу в програмі на структурному рівні

Всі терми, які трапляються в документах оброблюваної колекції, можна впорядкувати. Якщо тепер для деякого документа виписати по порядку ваги всіх термів, включаючи ті, яких немає в цьому документі, вийде вектор, який і буде представленням цього документа у векторному просторі. Розмірність цього вектора, як і розмірність простору, дорівнює кількості різних термів у всій колекції, і є однаковою для всіх документів.

Більш формально $$d_j = (w_1, w_{2j}, ..., w_{nj}),$$ (10)

де d_j – векторне уявлення *j*-го документа, w_{ij} – вага і- го терму в *j*- му документі, *n* – загальна кількість різних термів у всіх документах колекції [7-8].

Маючи таке подання для всіх документів, можна, наприклад, знаходити відстань між точками простору і тим самим вирішувати задачу подібності документів – чим ближче розташовані точки, тим більше схожі відповідні документи. У разі пошуку документа за запитом, запит теж представляється як вектор того ж простору, а отже, можна обчислювати відповідність документів запиту. Отже, підсумовуючи вищенаведені теоретичні викладки, будуємо такий алгоритм роботи підсистеми. Опис підсистеми аналізу тональності зображений нижче на рис. 1.

Під час роботи спочатку знаходимо результат за допомогою методу опорних векторів:

1. Розраховуємо за допомогою підходу TF-IDF вагу для кожного слова в словнику
2. Формуємо векторну модель для подальшого навчання SVM-класифікатора.
3. Будуємо гіперплощини, що класифікують слова відповідно до двох класів.
4. Обчислюємо гіпотезу.

Потім знаходимо результат за допомогою методу ключових слів:

1. Для кожного слова обчислюємо релевантну частоту (Relevance Frequency, RF).
2. Для кожного класу: обираємо підмножину слів з частотою більше за граничне значення цього класу.
3. Отримали список підмножин пар слів та їх частот, які відповідають кожному класу.
4. Кожному класу ставимо у відповідність суму частот по всім словам з відповідної підмножини.
5. Обираємо клас з найбільшою сумою частот, який буде гіпотезою.

Після виконання обох методів комбінуємо їх результати за певною стратегією. Сутність використання комбінованого методу полягає у постійному навчанні програми, а саме в аналізі текстової інформації на базі стартової вибірки та фінальне порівняння результатів роботи алгоритмів з відомими правильними результатами. Оцінки, які використовують, для порівняння результатів можна також використати для порівняння різних методів. Але перед цим розглянемо основні метрики.

Точність і якість системи аналізу тональності тексту оцінюється тим, наскільки добре вона узгоджується з думкою людини щодо емоційної оцінки досліджуваного тексту. Для цього можуть використовуватися такі метрики як точність та повнота [9]. Формула для знаходження повноти має вигляд:

R=correctly extracted opinions/total number of opinions,

де correctly extracted opinions – вірні думки; total number of opinions – загальна кількість думок (як знайдених системою, так і не знайдених). Точність обчислюється за формулою:

P = correctly extracted opinions\total number of opinions found by system,

де correctly extracted opinions – вірні певні думки, total number of opinions found by system – загальна кількість думок знайдених системою. Таким чином, точність виражає кількість досліджуваних текстів, пропозицій або документів, в оцінці яких думка системи аналізу тональності збіглася з думкою експерта.

При цьому на базі статті ”Автоматический анализ тональности текстов на основе методов машинного обучения” Котельников Е.В., Клековкина М.В. [10], експерти, зазвичай, погоджуються в оцінках тональності конкретного тексту в 79% випадків. Отже, програма, яка визначає тональність тексту за допомогою комбінованого методу з точністю 70%, робить це майже так само добре, як і людина. Цей результат може повністю влаштовувати нас, так як специфіка завдання полягає у відсутності точної відповіді. Ми можемо тільки намагатися видавати ті ж самі результати, що й отримувала людина, яка навчала нашу програму.

Висновки:

1. Проведено аналіз алгоритмів та методик, що здійснює автоматизацію процесу визначення тональності текстової інформації, які ґрунтуються на реалізації різноманітних способах, методах і алгоритмах та практичної спрямованості.

2. Проведено аналіз і обґрунтування вибору раціонального алгоритму аналізу тональності різномовної текстової інформації для задачі моніторингу інформаційного простору з метою виявлення джерел інформаційного впливу і запропоновано комбінований метод, який поєднує переваги методу SVM та методу ключових слів без отримання додаткових недоліків.

3. Визначено що комбінований метод автоматичного визначення тональності тексту, що об'єднує результати двох потужних класифікаторів (SVM та на основі ключових слів), дає можливість досягти підвищення якості класифікації в порівнянні не лише з цими класифікаторами, а й з найкращими результатами інших підходів.

ЛІТЕРАТУРА:

1. Терехов, А.В. Информатика: учеб. пособие / Терехов, А.В. Чернышов, В.Н. – Тамбов : Изд-во Тамб. гос. техн. ун-та, 2007. – Т35, С. 54-55.

2. L. Lytvynenko "Construction of premorphological level of analisys for multilanguage texts" // – USA. The advanced science journal, volume 2013 issue 5. – С. 28-32.

3. O. Nikolaievskyi "Components of lingware for automatic morphological analysis in knowledge-oriented machine translation system" // – USA. The advanced science journal, volume 2013 issue 5. – С. 32-36.

4. Угланова И.А., Ерофеева Е.В. Частотная категория в языке и речевой деятельности II. Слово отзовется: Памяти Аллы Соломоновны Штерн и Леонида Вольковича Сахарного. – Пермь: Перм. госуд. ун-т, 2006. С. 197-203.

5. Atkins S., Rundell M. The Oxford Guide to Practical Lexicography. Oxford: Oxford University Press, 2008. – 552 p.

6. Liu Z., Huang W., Zheng Y., Sun M. Automatic keyphrase extraction via topic decomposition. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, Massachusetts, 2010, pp. 366–376.

7. Пазельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: сб. научных статей. Вып. 10(17). М.: Изд-во РГГУ, 2011. С. 510–522.

8. Официальный сайт компании “Эр Си О”. Компонент определения тональности текста [Электронный ресурс] (http://x-file.su/tm/Default.aspx).

9. Ермаков А.Е., Киселев С.Л. Лингвистическая модель для компьютерного анализа тональности публикаций СМИ. Компьютерная лингвистика и интеллектуальные технологии: междунар. конф. Диалог'2005. М.: Наука, 2005.

10. Котельников Е. В., Клековкина М. В. Автоматический анализ тональности текстов на основе методов машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». Вып. 11 (18), М.: Изд-во РГГУ, 2012, С. 27-36.

REFERENCES:

1. Terekhov, A.V., Chernыshov, A.V. (2007). *Ynformatyka, ucheb. posobye*. [Informatics: studies. manual].Tambov, vol. 35, pp. 54-55 (In Russian).

2. Lytvynenko, L. (2013) Construction of premorphological level of analisys for multilanguage texts. The advanced science journal, vol. 2013, issue 5, pp. 28-32 (In USA).

3. Nikolaievskyi, O. (2013) Components of lingware for automatic morphological analysis in knowledge-oriented machine translation system. The advanced science journal, vol. 2013, issue 5, pp. 32-36 (In USA).

4. Ughlanova, Y.A., Erofeeva, E.V. (2006) *Chastotnaja kateghoryja v jazыke y rechevoj dejateljnosty II. Slovo otzovetsja: Pamjaty Allы Solomonovnы Shtern y Leonyda Voljkovycha Sakharnogho.* [Frequency category in a language and vocal activity]. Permj ghosud. un-t, pp. 197-203 (In Russian).

5. Atkins, S., Rundell, M. *The Oxford Guide to Practical Lexicography*. Oxford University Press, 2008, 552 p.

6. Liu, Z., Huang, W., Zheng, Y., Sun, M. (2010) Automatic keyphrase extraction via topic decomposition. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, Massachusetts, pp. 366–376.

7. Pazeljskaja, A.Gh., Solovj'ev, A.N. (2011) *Metod opredelenyja эmocyj v tekstakh na russkom jazыke.* [Method of determination of emotions in texts in Russian language]. Kompjjuternaja lynghvystyka y yntellektualjnыe tekhnologhyy: sb. nauchnыkh statej, issue. 10(17). Moscow, Publ.RGhGhU, pp.510–522 (In Russian).

8. Ofycyaljnыj sajt kompanyy "Эr Sy O". *Komponent opredelenyja tonaljnosty teksta*. [Component of determination of the key of text]. Available at: (http://x-file.su/tm/Default.aspx). (accessed 03.10.2017)

9. Ermakov, A.E., Kyselev, S.L. (2009) *Lynghvystycheskaja modelj dlja kompjjuternogho analyza tonaljnosty publykacyj SMY.* [Linguistic model for the computer analysis of the key of publications of MASS-MEDIA] Kompjjuternaja lynghvystyka y yntellektualjnыe tekhnologhyy. Moscow. (In Russian).

10. Koteljnykov, E.V., Klekovkyna, M.V. (2012) *Avtomatycheskyj analyz tonaljnosty tekstov na osnove metodov mashynnogho obuchenyja.* [Automatic analysis of the key of texts on the basis of methods of computer-aided instruction] Kompjjuternaja lynghvystyka y yntellektualjnыe tekhnologhyy: po materyalam ezheghodnoj mezhdunarodnoj konferencyy Dyalogh, Publ. RGhGhU, Moscow, pp. 27-36. (In Russian).

Рецензент: д.т.н., проф. Оксіюк О.Г., завідувач кафедри кібербезпеки та захисту інформації факультету інформаційних технологій Київського національного університету імені Тараса Шевченка

к.т.н., доц. Пампуха И.В., к.воен.н. Никифоров Н.Н., к.т.н. Лоза В.Н., Ступницкая О.И.

ОБОСНОВАНИЕ ВЫБОРА РАЦИОНАЛЬНОГО АЛГОРИТМА АНАЛИЗА ТОНАЛЬНОСТИ РАЗНОЯЗЫЧНОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ ДЛЯ ЗАДАЧИ МОНИТОРИНГА ИНФОРМАЦИОННОГО ПРОСТРАНСТВА

В работе проведен анализ алгоритмов и методик, что осуществляет автоматизацию процесса определения тональности текстовой информации. Проведено обоснование выбора рационального алгоритма анализа тональности разноязычной текстовой информации для задачи мониторинга информационного пространства с целью выявления источников информационного влияния, и предложен комбинированный метод, который сочетает преимущества метода SVM и методу ключевых слов без получения дополнительных недостатков. Определенно что комбинированный метод автоматического определения тональности текста, который объединяет результаты двух мощных классификаторов (SVM и на основе ключевых слов), дает возможность достичь повышения качества классификации в сравнении не только с этими классификаторами, но и с наилучшими результатами других подходов. Возникновение такого эффекта обеспечивается за счет учитывания степени уверенности классификаторов в своих решениях и подбора оптимальных параметров на основе скользящего контроля.

Ключевые слова: тональность, алгоритм, текстовая информация, мониторинг, информационное пространство.

Ph.D. Pampukha I.V., Ph.D. Nikiforov N.N., Ph.D. Loza V.N., Stupnitskaya O.I.

A GROUND OF CHOICE OF RATIONAL ALGORITHM OF ANALYSIS OF KEY OF POLYGLOT TEXT INFORMATION IS FOR TASK OF MONITORING OF INFORMATIVE SPACE

The analysis of algorithms and methodologies is in-process conducted, that carries out automation of process of determination of the key of text information. The ground of choice of rational algorithm of analysis of the key of polyglot text information is conducted for the task of monitoring of informative space with the aim of exposure of sources of informative influence and the combined method that combines advantages of method of SVM and method of keywords without the receipt of additional defects offers. Certainly that the combined method of autoplay of the key of text, that unites the results of two powerful classifiers (SVM and on the basis of keywords), gives an opportunity to attain upgrading of classification in comparing not only to these classifiers but also with the best results of other approaches. The origin of such effect is provided due to taking into account degrees of confidence of classifiers in the decisions and selection of optimal parameters on the basis of sliding control.

Keywords: key, algorithm, text information, monitoring, informative space.