УДК 32.973.202:07.681 к.т.н. **Марченко-Бабіч О.М.** (ВІКНУ)

ОСНОВНІ ВИМОГИ ДО СИСТЕМИ МОНІТОРИНГУ ІНШОМОВНИХ ЗМІ В ІНТЕРЕСАХ ЗАБЕЗПЕЧЕННЯ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ

У статті визначені основні функціональні вимоги до розроблення автоматизованої системи моніторингу іншомовних ЗМІ в інтересах забезпечення інформаційної безпеки. Береться до уваги фактор компетентності фахівців-розробників та визначаються основні етапи розробки. Приділяється увага особливостям організації мовної системи, зокрема рівням організації та структурі тексту. Визначається модель автоматизованого опрацювання тексту, що включає поєднання математичного апарату із оптимальним набором властивостей щодо аналізу тексту, у тому числі перевагами, важливими на даному етапі дослідження. Встановлюються елементи, що являють інтерес для розпізнавання при розмічуванні тексту. Запропоновано основні групи семантично-забарвлених лексичних одиниць для розпізнавання у тексті. Блок Сентимент-аналізу визначений основним інструментом оцінювання тексту на рівні розробки лінгвістичного забезпечення системи.

Ключові слова: природномовний, система, структура, рівні мовної системи, тезаурус, розпізнавання, Сентимент-аналіз.

Вступ. Сучасний інформаційний простір наповнений повідомленнями засобів масової інформації (ЗМІ), що висвітлюють події різних країн, у тому числі й України. Обсяги інформаційного контенту та існування його вже у стані безперервних потоків, які визначають стан інформаційного простору, вказують на те, що пошук необхідної інформації потребує не лише певних затрат часу, але, передусім, розробки відповідних спеціалізованих пошукових систем різної спрямованості.

Оскільки висвітлення подій про Україну в іноземних ЗМІ зазвичай здійснюється іноземними мовами, то опрацювання та аналіз інформаційних повідомлень вимагає значних витрат часу, ускладнюючи процес виявлення інформаційних джерел та своєчасне реагування на них. Тому вкрай важливо забезпечення процесу моніторингу іноземних ЗМІ сучасними програмно-математичними засобами для автоматизованого виявлення інформаційних джерел у серед таких ЗМІ та своєчасного реагування.

Аналіз останніх досліджень. В ході досліджень за даним напрямом останнім часом приділяється увага удосконаленню засобів оперативного й поглибленого аналізу інформаційних масивів, у тому числі текстових повідомлень ЗМІ. Так, це може досягатись шляхом автоматизації процесу багатовимірної аналітичної обробки повідомлень ЗМІ, що включає інтеграцію процесу обробки даних та динамічної актуалізації вихідних умов з відповідних інформаційних джерел [1]. Для аналізу інформаційного контенту і прогнозування його розвитку в Інтернет-просторі у [2] запропоновано інструментарій, що поєднує пошук релевантних джерел, аналіз вибраного контенту, прогноз його розвитку, і складається з математичних методик та технологічних компонувань даних у єдиний профіль для конкретної галузі застосування за напрямом інформаційної безпеки. Про зростаючу важливість моніторингу та контролю відповідними державними структурами різних сфер суспільно-політичного життя свідчить і те, що на даний час проводиться розробка системи виявлення терористичної загрози в Інтернет-комунікаціях з боку деяких країн в інтересах структур безпеки Європейського союзу [3].

Проте, незважаючи на значний науковий доробок у даній сфері, аналіз проблемних питань в автоматизованому опрацюванні інформаційних повідомлень [4] свідчить, що однією з найслабших ланок є процедура побудови формальної моделі їх семантики. Також недостатня увага приділялась розробці програмно-математичних засобів опрацювання зарубіжних іншомовних ЗМІ в інтересах інформаційної безпеки України з приділенням уваги лінгвістичній складовій цього процесу.

Тому **метою і основним змістом статті є** визначення основних вимог до системи автоматизованого моніторингу зарубіжних іншомовних ЗМІ в інтересах інформаційної

безпеки України, з урахуванням особливостей як лінгвістичного компоненту, так і воєнно-політичної обстановки саме для нашої держави.

Викладення основного матеріалу. Зазвичай основним методом моніторингу в центральних органах виконавчої влади є пошук інформації в Інтернет за напрямом діяльності за допомогою ключових слів. При цьому значний інтерес для опрацювання інформації становить аналітико-прогностичний напрям, виконання завдань за яким дозволяє сприяти управлінню державними процесами та є необхідним у діяльності інформаційно-аналітичних підрозділів державних структур, у тому числі Міністерства оборони України [5].

Тому в інтересах інформаційної безпеки України необхідним є створення вітчизняних засобів, які повинні здійснювати моніторинг текстів іншомовних ЗМІ щодо подій в Україні та виконувати аналітико-прогностичні завдання.

Основні функції системи моніторингу іншомовних ЗМІ повинні бути такими:

- збір в Інтернет іншомовних повідомлень іноземних ЗМІ про події в Україні;
- їх аналіз.

Система повинна враховувати сучасні досягнення в галузі кібернетики, комп'ютерної лінгвістики, психолінгвістики та психології. У розробці необхідна також участь фахівців-лінгвістів, які повинні володіти іноземними мовами, що є важливим інструментом у даному дослідженні, на рівні, достатньому для здійснення аналізу тексту на предмет відповідності його окремих компонентів умовам дослідження. Результати їх роботи повинні бути включені до бази знань підсистеми визначення емоційної тональності та підсистеми прогнозування реакцій аудиторії, які повинні увійти до системи моніторингу. Відповідна база знань повинна бути покладена в основу блоку Сентимент-аналізу (Sentiment-analysis) системи та безпосередньо впливати на прийняття нею рішення щодо змісту повідомлення.

Створення системи передбачає два основні етапи розробки:

- розробка лінгвістичного забезпечення;
- розробка програмного забезпечення.

Етап розробки лінгвістичного забезпечення включає аналіз природномовних текстів. Він повинен бути спрямований на формування понятєвої структури, тобто на автоматичний витяг знань з іншомовних текстів та їх прагматичну інтерпретацію в термінах поставленої задачі. При цьому кожен текст розглядатиметься як об'єкт різних рівнів аналізу: як знакова система, як граматична система та як система знань про предметну галузь. Через те, що кожен рівень має свої особливості, свої засоби вираження, він передбачатиме наявність відповідних методів обробки.

Лінгвістичне розпізнавання знань з предметної галузі відображатиме морфологічний, синтаксичний та семантичний рівні мовної системи. На лінгвістичному етапі розроблення системи слід побудувати поняттєву структуру тексту. Вона включатиме тезаурус, структура якого орієнтована на завдання дослідження. Результати графемного та синтаксичного розпізнавання є вхідними даними для семантичного розпізнавання, як і еталонні моделі з тезаурусом понять, тезаурусом відношень і тезаурусом логіко-семантичних відношень. На етапі семантичного розпізнавання всі фрагменти тексту повинні бути об'єднані в єдину логіко-семантичну структуру, обробка полягатиме в узагальненні та уніфікації понять, відношень та їх характеристик. Сутність процесу становить виділення головної ядерної структури, що відображає, про що йдеться у тексті. Структуру можна формалізовано представити у вигляді ядерного ланцюга:

S (суб′єкт)⟶A (дія)⟶O (об'єкт)

Така ядерна структура є універсальною і має багато різних інтерпретацій ("хто про що", "хто що робить" тощо). Ядерна структура повинна супроводжуватись додатковою інформацією про кожен її елемент. Цю модель слід наповнювати безпосередньо з тексту.

Заключною на цьому етапі є процедура семантико-прагматичного розпізнавання [6].

Її завдання передбачає інтегрування поняттєвої структури тексту до бази знань. Вхідними даними для цієї процедури є тезауруси з предметної області та результати

семантичного розпізнавання. Для формування бази знань необхідно створити текстовий корпус з розмічених іншомовних текстів.

В основу функціонування системи на даному етапі пропонується покласти Модель Байєса, зокрема простий імовірнісний класифікатор, заснований на застосуванні Теореми Байєса зі строгими (наївними) припущеннями про незалежність − так званий Наївний байєсовський класифікатор [8]. Його використання має такі переваги:

- здатність до ефективного навчання;
- мала кількість даних для навчання, необхідних для оцінки параметрів, що потрібні для класифікації.

Моделі Байєса повинна бути надана навчальна вибірка, за якою модель навчатиметься, та розраховуються вагові коефіцієнти (вірогідності для слів). Після цього модель, що навчається, стане здатною аналізувати вхідні тексти.

Процес аналізу тексту передбачатиме подачу текстового документу на вхід програми. На виході повинен бути виданий опрацьований текст із розміченими лексичними одиницями, де визначено, який сегмент тексту містить емоційний заряд і спонукає до відповідного реагування. 2 основні фази існування програми включатимуть:

- фазу навчання;
- фазу експлуатації.

В ході навчання програми повинно проводитись обчислення вагових коефіцієнтів по розмічених зразках. В ході експлуатації програма, що вже навчилась, повинна знаходити лексичні одиниці з певним емоційним навантаженням, оцінювати його та аналізувати.

На початковому етапі для визначення програмі завдань з розмічування тексту слід встановити, які саме його елементи становлять інтерес для розпізнавання. На даному етапі досліджень передбачається їх представлення двома групами семантично-забарвлених лексичних одиниць: першої − з лексичних одиниць, що формують емоційну тональність тексту; другої – з відповідників, що мають відношення до функціональних психічних станів людини.

Вищезгадана лексика повинна опрацьовуватись за допомогою сентимент-аналізу. Сентимент-аналіз (Sentiment-analysis) − це область комп'ютерної лінгвістики, що займається вивченням думок і емоцій у текстових документах і являє собою сукупність методів, розрахованих на автоматичне виявлення емоційної оцінки (тональності, або сентименту), вираженої в тексті. Такий аналіз дозволяє охарактеризувати емоційне забарвлення тексту − позитивне, негативне або нейтральне, виявити суб'єкт і об'єкт цього тексту [9].

Виходячи з того, що до блоку сентимент-аналізу повинна входити лексика, яка позначає властиві психіці людини явища, то у даному випадку інтерес представляють особливості функціонування психіки при певних зовнішніх впливах.

Висновки. В сучасних умовах безпеки опрацювання іншомовних ЗМІ повинно включати функції аналізу та прогнозування. Розробка відповідної системи передбачає роботу над створенням блоку лінгвістичного забезпечення з подальшим створенням блоку програмного забезпечення.

Робота над лінгвістичним забезпеченням передбачає аналіз природномовних текстів із відображенням морфологічного, синтаксичного, семантичного та прагматичного рівнів мовної системи. Дані, отримані з аналізу текстів, повинні бути покладені в основу бази знань, яку формує текстовий корпус з розмічених іншомовних текстів.

Однією з найбільш відповідних моделей для функціонування системи є Модель Байєса, яка здатна до швидкого навчання та не потребує великої кількості набору ознак для прийняття рішення. Інструментом опрацювання лексики бази знань варто вибрати сентимент-аналіз, що дозволить врахувати особливості функціонування психіки людини при зовнішніх впливах на неї.

Усе це покликано забезпечити оперативне виявлення інформаційних джерел серед іноземних ЗМІ та своєчасне реагування.

ЛІТЕРАТУРА:

1. Писарчук О.О. Технологія автоматизованої багатовимірної оперативної та поглибленої аналітичної обробки актуальних інформаційних масивів / О.О. Писарчук, В.С. Косіков // Проблеми створення, випробування, застосування та експлуатації складних інформаційних систем. – 2015. – Вип. 10. –С. 183–195. – Режим доступу: http://nbuv.gov.ua/UJRN/Psvz_2015_10_22

2. Писарчук О.О. Методика прогнозування розвитку інформаційного контенту в мережі Інтернет / О. О. Писарчук, Д. В. Порада // Проблеми створення, випробування, застосування та експлуатації складних інформаційних систем. – 2015. – Вип. 10. – С. 170–182. – Режим доступу: http://nbuv.gov.ua/UJRN/Psvz_2015_10_21.

3. Dr. Babych B., Dr.Atwell E. Multilingual information extraction framework for real-time detection of terrorist propaganda threats in on-line communication. Військова освіта і наука: сьогодення та майбутнє: збірник матеріалів XI Міжнародної науково-практичної конференції. 27 листопада 2015 р. – К. : 2015. – С.93.

4. Проблеми автоматизованого аналізу і оброблення природномовних текстів. Бойко О.В., Мірошниченко О.В., Стамбірська Р.Г. Всеукраїнська науково-практична конференція молодих вчених, ад'юнктів, слухачів, курсантів і студентів. 24 квітня 2015 р. – К., 2015. – ВІКНУ. – С.243.

5. Г.В. Любовець. Ситуаційний сервіс Міністерства оборони України як намагання держави протистояти новим загрозам геополітичного та техногенного характеру // Любовець Г.В., Король В.Г. Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К. : ВІКНУ, 2015. – Вип. № 50. – С. 112—118.

6. Бадьорина Л.М. Основи комп'ютерної лінгвістики. Навчальний посібник / Л.М.Бадьорина, І.В.Замаруєва, В.А.Широков. – К. :Видавничий центр КНУКІМ, 2011. – С.107-108.

7. Андреев А.М. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа. А.М. Андреев, Д.В.Березкин, Морозов В.В., Симаков К.В. НПЦ «ИНТЕЛТЕК ПЛЮС». – Режим доступу:
http://www.inteltec.ru/publish/articles/textan/RCDL2003.shtml

8. Зверева П.П. Сентимент-анализ текста (на материале печатных текстов газеты "The New York times" "О России и россиянах"). – М. : Вестник МГОУ, Серия: Лингвистика, № 5, 2014. – Режим доступу: http://vestnik-mgou.ru/Articles/Doc/7774.

REFERENCE:

1. Pisarchuk O.O. Tekhnologiya avtomatizovanoї bagatovymirnoї operatyvnoї ta pogliblenoї analitychnoї obrobki actualnych informatsyiykh masyviv / O.O. Pisarchuk, V.S. Kosikov // Problemy stvorennya, vyprobuvannya, zastosuvannia ta ekspluatatsiyi skladnynykh informatsyinykh system. – 2015. – Iss. 10. – Page 183–195. – Mode to access: http://nbuv.gov.ua/UJRN/Psvz_2015_10_22

2. Pisarchuk O.O. Metodyka prognozuvannia rozvytku informatsiynogo contentu v meregi internet / O.O. Pisarchuk, D.V. Porada // Problemy stvorennya, vyprobuvannia, zastosuvannia ta ekspluatatsiyi skladnykh informatsyinykh system. – 2015. – Iss. 10. – Page 170–182. – Mode to access: http://nbuv.gov.ua/UJRN/Psvz_2015_10_21

3. Multilingual information extraction framework for real-time detection of terrorist propaganda threats in on-line communication: materialy XI Mizhnarodnoї naukovo-practychnoyї conferentsiї, [Viyskova osvita i nauka: siogodennia ta maybutne], (Kyiv, 27 lystopada 2015r.) / K. : VIKNU, 2015 — Page 93.

4. Problemy avtomatizovanogo analizu i obroblennia pryrodnomovnykh tekstiv. Boyko O.V., Miroshnichenko O. V., Stambirska R. G. Vseukrainska naukovo-practychna conferentsiya molodykh vchenikh, ad′yunktiv, slukhachiv, cursantiv i studentiv. 24 04. 2015 r. K. : 2015. VIKNU. — Page 243.

5. G.V. Lyubovets. Situatsiyniy service Ministerstva oborony Ukrany yak namagannya dergavy protistoyaty novym zagrozam geopolitychnogo ta technogennogo characteru // Lyubovets G.V. , Korol V. G. Zbirnyk naukovykh prats Viyskovogo institutu Kyïvskogo natsionalnogo universitetu imeni Tarasa Shevchenka. – K : VIKNU, 2015. – Iss. №. 50. – Page 112–118.

6. Badyorina L.M. Osnovy komp'yuternoї lingvistiki. Navchalnyi posibnyk / L. M. Badyorina, I. V. Zamaruyeva, V. A. Shirokov. – K. : Vydavnychyi center KNUKIM, 2011. Page 107 – 108.

7. Andreyev A.M. Avtomatycheskaya classificatsiya tekstovykh documentiv s ispolzovaniem neyrosetevykh algorytmov i semanticheskogo analizu. A.M. Andreyev, D.V.Berezkin, V. V.Morozov , K.V.Simakov . NPC "INTELTEKH PLUS". Mode to access: http://www.inteltec.ru/publish/articles/textan/RCDL2003.shtml

8. Zvereva P.P. Sentiment-analiz texta (na materiale pechatnykh tekstiv gazety "The New York Times" "About Russia and Russians"). — M. : Vestnik MGOU, Seriya: Linguistica, №. 5, 2014. Mode to access: http://vestnik-mgou.ru/Articles/Doc/7774

Рецензент: д.т.н., проф. Замаруєва І.В., Державний університет телекомунікацій

к.т.н. Марченко-Бабич О.Н.

ОСНОВНЫЕ ТРЕБОВАНИЯ К СИСТЕМЕ МОНИТОРИНГА ИНОЯЗЫЧНЫХ СМИ В ИНТЕРЕСАХ ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

В статье определены основные функциональные требования к разработке автоматизированной системы мониторинга иноязычных СМИ в интересах обеспечения информационной безопасности. Принимается во внимание фактор компетентности специалистов-разработчиков и определяются основные этапы разработки. Уделяется внимание особенностям организации языковой системы, в частности уровням организации и структуре текста. Определяется модель автоматизированной обработки текста, которая включает объединение математического аппарата с оптимальным набором свойств относительно анализа текста, в том числе преимуществами, важными на данном этапе исследования. Устанавливаются элементы, которые представляют интерес для распознавания при разметке текста. Предложены основные группы семантически-окрашенных лексических единиц для распознавания в тексте. Блок Сентимент-анализа определен как основной инструмент оценивания текста на уровне разработки лингвистического обеспечения системы.

Ключевые слова: естественноязыковой, система, структура, уровни языковой системы, тезаурус, распознавание, Сентимент-анализ

Ph.D. Marchenko-Babich O.N.

THE BASIC REQUIREMENTS FOR THE FOREIGN LANGUAGE MASS MEDIA MONITORING SYSTEM FOR THE PURPOSE OF THE INFORMATION SECURITY

The article defines the basic requirements for the foreign language mass media automated monitoring system for the purpose of the information security. The factor of professional competence of the system developers is taken into consideration, and the main phases of development are defined. The attention for the features of the language system organisation is paid particularly for the main levels of the text structure and organisation. The model of the automated text processing is defined. It includes mathematical apparatus combined with text analysis optimum signs set. Besides, advantages considerable for the current stage of research are taken into account. Elements important for the indication in the process of text marking are set. Basic groups of semantically coloured lexical units are proposed for their identification in the text. The Sentiment Analysis block is set as the principle tool for the text evaluation on the level of the system linguistic support.

Key words: in natural language, system, structure, language system levels, thesaurus, identification, Sentiment-analysis.