УДК 32.973.202:07.681

д. ф.-м.н., доц. **Марченко О.О.** (КНУ)
к.т.н. **Марченко-Бабіч О.М.** (ВІКНУ)

РОЗРОБКА ЛІНГВІСТИЧНОГО ЗАБЕЗПЕЧЕННЯ ДЛЯ АВТОМАТИЗОВАНОГО МОНІТОРИНГУ СОЦІАЛЬНИХ МЕРЕЖ

У статті розглядаються нові принципи та підходи до розробки лінгвістичного забезпечення для автоматизованого семантичного пошуку та відбору інформації, зміст якої може свідчити про цілеспрямований вплив на аудиторію, серед текстів соціальних мереж та інших соціально-орієнтованих Інтернет-сервісів. Визначено особливості мовних конструкцій та компонентів бази знань системи автоматизованого моніторингу. Запропоновано спосіб підвищення точності смислового аналізу текстів, який включає уникнення впливу полісемії багатозначних слів та словосполучень. Підходи, визначені у статті, дають можливість отримати в процесі пошуку текст, який не містить жодного ключового слова із запиту і при цьому повністю відповідає йому за змістом та семантикою. Пропонуються види аналізу тексту, які доцільно використовувати для семантичного пошуку в текстах соціальних мереж та інших соціально-орієнтованих Інтернет-сервісах: контекстний асоціативно-семантичний та Sentiment Analysis, що дає можливість врахування багатьох нюансів та деталей емоційного забарвлення текстових повідомлень і може відображати загальну спрямованість інформаційного впливу, що здійснюється зацікавленими учасниками інформаційного простору.

Ключові слова: інформаційний вплив, інформаційний простір, соціальна мережа, Інтернет-сервіс, обробка природної мови, алгоритм пошуку.

Вступ. Останнім часом спостерігається швидкий темп розвитку, розповсюдження та запровадження в усі сфери життя глобальної інформаційної мережі Інтернет. Значне зростання обсягу та швидкості передачі інформації зумовило розвиток діяльності іноземних

спеціальних структур, відповідальних за здійснення психологічного тиску на військовослужбовців та населення інших країн. Країни, що претендують на глобальне лідерство (або на його повернення), постійно нарощують свої потужності зі здійснення інформаційного впливу. Серед таких країн визначаються як загальновизнані світові лідери (як, наприклад, США), так і ті, що останнім часом намагаються повернути собі втрачені сфери впливу (як, наприклад, Російська Федерація).

Сучасні спеціалісти поряд з такими традиційними методами як друкована пропаганда, усна агітація, теле- і радіомовлення, в інтересах ІВ стали активно використовувати мережу Інтернет, зокрема, технології соціальних мереж та інших соціально-орієнтованих Інтернет-сервісів [1].

За висновками фахівців у галузі інформаційних технологій, найбільша загроза монополізації інформаційного впливу на даний час походить від Google, YouTube, Facebook, Twitter, «ВКонтакте» та інших крупних ресурсів. Зростає покоління людей, для яких Інтернет невідділимий від соціальної мережі, в якій вони сидять, і саме з якої вони отримують велику частку інформації [2].

Кардинально змінилась роль «простих громадян» у формуванні інформаційного простору, який створюється засобами соціальних медіамереж: кожен з них стає не лише об’єктом, а й ланкою, що приймає інформаційний вплив та сприяє його подальшому поширенню. Дослідження [3] свідчать, що ці мережі та інші соціально-орієнтовані Інтернет-сервіси можуть використовуватись як інструмент маніпулювання суспільною (масовою, індивідуальною) свідомістю та є зручним майданчиком для формування громадської думки.

Європейська дослідницька компанія InsitesConsulting.eu підрахувала, що різними соціальними мережами в усьому світі зараз користуються більш ніж 1 млрд людей. Останнім часом до соціальних мереж приєдналось більш ніж 70% усіх Інтернет-користувачів [4]. Хвиля "facebook"-революцій в різних кутках світу в останні роки дійсно продемонструвала здатність Інтернету грати провідну роль серед ЗМІ, оскільки це наймасовіший, найдешевший та найважче контрольований державою засіб масової інформації. Усе це створює необхідність запровадження інструментів відстеження та контролю інформаційного потоку соціальних ресурсів з метою відповідного реагування.

Усе вищевикладене вказує на необхідність удосконалення технологій пошуку та опрацювання повідомлень соціальних мереж з ознаками інформаційного впливу за допомогою програмних засобів автоматизованого моніторингу. Актуальним є винайдення нових підходів та вдосконалення вже існуючих у розробці лінгвістичного забезпечення цих засобів та побудова алгоритмів обробки природномовних текстів.

Аналіз останніх досліджень В ході проведених досліджень у даній галузі вже визначено інформаційно-комунікативні аспекти, встановлено психолінгвістичні особливості текстів соціальних мереж та інших соціально-орієнтованих Інтернет-сервісів [5]. Також досліджено механізм здійснення інформаційного впливу засобами цих Інтернет-сервісів, розглянуто основні риси їх інформаційного контенту [3] із зазначенням необхідності його моніторингу та контролю відповідними державними структурами. Для структурування контенту соціальних ресурсів Інтернет розроблено процедуру фільтрації інформаційних повідомлень [6]. Разом з тим**,** для реалізації зазначених заходів не визначено параметрів формування лінгвістичного забезпечення автоматизованого пошуку та відбору повідомлень, що містять задану змістовно-емоційну складову, у вказаному сегменту Інтернету.

Тому **метою і основним змістом статті** є визначення принципів та підходів до побудови лінгвістичного забезпечення для автоматизованого семантичного пошуку та відбору інформації, зміст якої може свідчити про цілеспрямований вплив на аудиторію, серед текстів соціальних мереж та інших соціально-орієнтованих Інтернет-сервісів.

Викладення основного матеріалу. Розробка лінгвістичного забезпечення засобів автоматизованого семантичного пошуку включає визначення підходів для забезпечення найбільш якісного аналізу та інтелектуальної обробки об’єкта пошуку. Так, останнім часом серед найбільш актуальних напрямів виділився такий помітний сегмент ІТ-досліджень, як

новий підхід, що має інтегрувати в собі риси та характеристики різних технологій для побудови та аналізу об'ємної багатовимірної моделі інформаційних об'єктів. Побудована модель за допомогою багаторівневого представлення різної за походженням та модальністю інформації про об'єкт виводить якість аналізу та інтелектуальної обробки текстів на суттєво вищий рівень, з огляду на наявність різнопланових точок зору на інформаційні об'єкти вхідного тексту.

Одним з компонентів такої багатовимірної моделі є мовна складова, яка передбачає врахування особливостей дискурсу соціальних мереж. Його важливою властивістю є форма міжособистісного та міжгрупового спілкування, яка відрізняється зниженим рівнем критичності та потреби у верифікації інформації.

Особливостями дискурсу соціальних мереж та інших соціально-орієнтованих Інтернет-сервіісів є [5]:

- конкретність та образність ключових слів у дискурсі офіційних і неофіційних Інтернет-ресурсів і посиланнях пошукових систем;
- емоційне перенасичення тексту: метафори та інші образні засоби, що підмінюють фактичний матеріал;
- використання риторичних запитань, що підштовхують читача до потрібних відповідей;
- приховування джерел інформації (з посиланням на "деяких експертів" тощо);
- використання лінгвістичних структур єдності, довіри;
- експлуатація ідеї "кола своїх", навмисне включення до нього мережевого ресурсу;
- використання евфемізмів, що формує необхідний маніпулятору образ;
- властивостями слова є стискання, розширення та злиття з іншими словами;
- візуальне підкріплення змісту переданої інформації, оскільки немовні моменти спілкування менше піддаються осмисленому контролю: "ні з чого" виникає певне емоційне ставлення.

Крім того, тексти соціальних мереж присвячені актуальним на визначений період темам і за своїми лінгвістичними характеристиками відображають особливості Інтернет-спільноти, залученої до спілкування. Це знайшло своє відображення й у керівних документах з інформаційного забезпечення для фахівців з інформаційних операцій США [7], залучених до роботи зі ЗМІ, у тому числі Інтернет. Так, правила такої роботи включають:

- створення інформаційного простору з таким змістовно-емоційним наповненням, що забезпечить слідування меті інформаційної операції;
- обговорення змісту, що забезпечує дотримання ключових тем, які необхідно дотримуватись або навпаки, яких слід уникати;
- дотримання у повідомленнях стилю викладення, що відображає соціокультурні особливості цільової аудиторії та лінгвістичні особливості середовища, на яке спрямовано вплив.

Виконання цих правил забезпечується наданням визначеним підрозділам відомостей щодо особливостей цільової аудиторії та інших складових інформаційного середовища.

З огляду на це, лінгвістична складова системи автоматизованого моніторингу повинна враховувати мовні конструкції, властиві дискурсу соціальних мереж. Вона також повинна відображати загальні тенденції розвитку у сфері, якої стосуються повідомлення зазначеного Інтернет-сегменту. Цю базу слід періодично оновлювати та настроювати згідно поточного інформаційного контенту.

До бази знань даної системи слід включити такі складові:

визначений профіль пошуку (розділи новин в електронних ЗМІ, певні сайти, що є місцем спілкування Інтернет-спільнот);

тематику, визначену напрямом діяльності;

емоційну забарвленість текстів повідомлень;

мову повідомлення, визначену завданням пошуку.

Одним із засобів, здатних встановити емоційну забарвленість тексту, є технологія Sentiment analysis, що дозволяє розподілити повідомлення за характером на позитивні та негативні згідно оціночних суджень їх авторів про предмет обговорення. Завдяки фільтрам на інформацію певного характеру, наприклад, негативну, є можливість відбирати тексти певної спрямованості згідно завдань пошуку.

При здійсненні моніторингу необхідно враховувати пошуковий профіль користувача (із врахуванням особливостей спектру його інтересів), а також можливість завдання пошуку не лише запитами, але й прикладами еталонних документів за їх «образом і подобою» за змістом і семантикою. Новизна підходу полягає у тому, що документ шукається не за принципом співпадіння ключових слів, а за принципом відповідності семантичних структур знайденого документу пошуковому запиту. Саме завдяки цьому вдається долати проблеми погіршення точності смислового аналізу текстів внаслідок впливу полісемії багатозначних слів та словосполучень [8].

Крім того, пропонується запровадження алгоритмів семантичного пошуку, які допомагають поширювати інформаційно-пошукові запити за допомогою синонімів, семантично-близьких понять (термів), що містяться в семантичній базі знань системи [9]. Це дасть змогу формалізувати процес складання ефективного пошукового запиту, побудувати синонімічний ряд для кожного зі слів та вкласти до пошукової системи усі необхідні дані. Таким чином можна знайти такий текст, який не містить жодного ключового слова з запиту і при цьому повністю за змістом та семантикою відповідає даному запиту.

Для знаходження повідомлень за вказаним напрямом проводиться лінгвістичний аналіз, складовими якого є лексико-морфологічний, синтаксичний, семантичний аналіз для отримання певної семантичної структури, яку можна проаналізувати з точки зору впливу на цільову аудиторію. Згідно з синтаксичною структурою речень текстів будуються семантичні графи та проводиться психолінгвістичний аналіз їх компонентів.

Особливостями повідомлень, якими обмінюються учасники Інтернет-спільноти, є те, що вони часто представлені у вигляді коротких текстів (наприклад «твітів»), які не піддаються стандартним алгоритмам. Для відстеження даних повідомлень необхідно використання алгоритмів, спеціально пристосованих для обробки таких текстів. Тому з урахуванням нових особливостей Інтернет-контенту з'являється все більше спеціалізованих пошукових систем, які використовуються для пошуку на сайтах з конкретної тематики. Згідно останніх наукових досліджень [10], при вирішенні зазначених завдань для побудови бази знань доцільно використовувати як тексти довільної форми, так і напівструктуровані джерела інформації (таблиці, списки, сайти регулярної структури). Також слід приділити увагу системам безперервного навчання, наприклад такій, що реалізована у проекті NELL [11] та ітераційно виконує дві задачі: задачу читання і задачу навчання. Під задачею читання розуміється отримання системою нових фактів з неструктурованих або напівструктурованих джерел (текстів). Задача навчання — на отриманих фактах сформувати нові патерни для більш ефективного «читання» системою текстових масивів мережі Інтернет [11].

Автоматичний розподіл відібраних повідомлень доцільно здійснювати залежно від особливостей висвітлення у них об’єкта пошуку. Для цього у процесі семантичного аналізу повідомлень соціальних мереж пропонується використання підходу, що застосовується у системах семантичного моніторингу [12]. Зокрема, це контекстний асоціативно-семантичний аналіз для обробки текстових потоків і корпусів з блоком якісного оцінювання лінгвістичних фокусних об’єктів. Він дозволяє обчислювати якісні характеристики й параметри будь-якого заданого лінгвістичного об’єкта в корпусах текстів і текстових потоках, відстежуючи динаміку змін та визначаючи основні тенденції оцінювання фокусного об’єкта. Після подачі на вхід системи імені заданого об’єкта вона формує семантичний фокус-образ в мережі онтології, обчислюючи якісні характеристики і параметри заданого об’єкту в тексті. Важливим етапом створення системи семантичного моніторингу є формування лінгвістичної шкали для якісних оціночних концептів онтології. Перший підхід визначення чисельно-порядкових значень концептів виконується за допомогою асоціативно-контекстних

алгоритмів, які шукають відстані в мережі онтології між поточним концептом і концептом-максимумом (мінімумом) даної шкали. Другий підхід, задіяний при розробці лінгвістичної шкали, використовує частотні алгоритми, що визначають частоту спільної появи пар слів в глобальних корпусах текстів, встановлюючи таким чином близькість їх семантичних значень (із врахуванням винятків серед сполучень певних груп слів). Такий контекстний асоціативно-семантичний аналіз дозволяє гнучко варіювати значення якісних оціночних концептів залежно від локально-глобального контексту, що дає можливість враховувати складні з точки зору ординарної семантики випадки застосування лексики.

Використання такого підходу забезпечує обчислення якісних характеристик і параметрів тексту з відстеженням динаміки змін та визначенням основних тенденцій оцінювання об'єкта вивчення. В свою чергу, застосування підходів напряму Sentiment Analysis [13] в системах моніторингу та пошуку текстів соціальних медіамереж дає можливість врахування багатьох нюансів та деталей емоційного забарвлення текстових повідомлень, що є дуже затребуваним з огляду на специфіку інформаційного простору, який формується такими повідомленнями та може відображати загальну спрямованість інформаційного впливу, що здійснюється зацікавленими учасниками інформаційного простору.

Висновки. З огляду на швидкий розвиток Інтернет-простору, який вже сьогодні є наймасовішим, найдешевшим та найважче контрольованим державою засобом масової інформації, необхідно запровадження інструментів відстеження та контролю його інформаційного потоку. Це, зокрема, соціальні мережі та інші соціально-орієнтовані Інтернет-сервіси, у формуванні інформаційного простору яких зацікавлені певні сили. Тому для оперативного відстеження інформаційного потоку всередині цих соціально-орієнтованих мереж необхідно залучення програмних засобів автоматизованого моніторингу, лінгвістичне забезпечення яких повинно включати:

- базу знань, яка враховує особливості дискурсу повідомлень, властивих соціальним мережам, профілю пошуку, об'єктів пріоритетного вивчення Інтернет-повідомлень. Цю базу слід періодично оновлювати та настроювати згідно поточного інформаційного контенту;
- алгоритми як для обробки неструктурованих даних (звичайних текстів новин тощо), так і для напівструктурованих даних (таблиць, списків, сайтів регулярної структури) та коротких повідомлень.

Основні етапи обробки текстових повідомлень соціально-орієнтованих Інтернет-сервісів, у тому числі соціальних мереж, повинні включати:

формалізацію повідомлень, що передбачає побудову семантичних графів згідно із синтаксичною структурою речень в текстах повідомлень та подальшим психолінгвістичним аналізом компонентів графу;

автоматичний розподіл повідомлень, відібраних з мережі Інтернет програмними засобами, з урахуванням актуальності повідомлення та характеристик джерел, які їх поширюють, а також особливостей висвітлення об'єктів, що становлять інтерес.

Важливо також врахування особливостей пошукового профілю користувача із можливістю завдання пошуку не лише запитами, але й прикладами еталонних документів за їх «образом і подобою», не за принципом простого співпадіння ключових слів, а за принципом відповідності семантичних структур знайденого документу запиту користувача, із застосуванням підходів напряму Sentiment Analysis.

Запропоновані принципи побудови лінгвістичного забезпечення дають можливість оперативного пошуку та відбору інформації, зміст якої може свідчити про активність зацікавлених структур відносно української аудиторії, серед текстів соціальних мереж в Інтернет, з метою своєчасного адекватного реагування.

ЛІТЕРАТУРА:

1. Китов П. Совершенствование способов и средств ведения психологических операций вооружённых сил США /П. Китов // Зарубежное военное обозрение. – М. : «Издательский дом «Красная звезда», 2013. – №3 (792), г. Москва. – С. 19–22 .

2. Александр Ольшанский: как Украине победить в информационной войне с РФ и чем опасны Google, Twitter и Facebook. Режим доступу до сайту: http://ain.ua/2015/04/21/576406.

3. Панченко В.М. Інформаційна безпека особи в умовах соціалізації Інтернет-сервісів. Актуальні проблеми управління інформаційною безпекою держави: зб. матер. наук.-практ. конф., 30 березня 2012. – К.: Наук-вид. відділ НА СБ України, 2012. – С. 81-83.

4. Филонов Д. Более миллиарда человек пользуются социальными сетями во всем мире. Режим доступу до сайту: http://digit.ru/internet/20110914/384078152.html

5. Сугестивні технології маніпулятивного впливу: навчальний посібник/ [В. М. Петрик, М. М. Присяжнюк, Л. Ф. Компанцева, Є. Д. Скулиш, О. Д .Бойко, В. В. Остроухов]; за заг. ред. Є. Д. Скулиша − К. : Науково-видавничий відділ НА СБ України, 2010 − 248 с.

6. Писарчук О.О. Методика фільтрації інформаційних повідомлень Інтернет-джерел та їх класифікація. / Писарчук О.О., Стиров Н.В.// Проблеми створення, випробування, застосування та експлуатації складних інформаційних систем. Збірник наукових праць ЖВІ ДУТ.– 2014. − Вип.9.− С. 49–55.

7. JP 3-13.2. Military Information Support Operations. Joint Chiefs of staff, 07 January 2010 Incorporating Change 1 20 December 2011, V-3.

8. Марченко О. О. Моделювання семантичного контексту при аналізі текстів на природній мові. Вісник Київського університету. Сер. фіз.-мат. науки, 2006. – № 3. – С. 230–234.

9. Анісімов А.В., Марченко О.О., Никоненко А.О. UWN: Універсальна онтологічна база знань украньскої мови . Проблеми програмування, 2012, № 2-3 – С. 348–355.

10. Глибовець А.М. Алгоритми обробки текстів вільної форми для отримання фактів і зв'язків між ними / Глибовець А.М., Марченко О.О., Циганок Д.В., Бабіч О.М. // Наукові записки НаУКМА. Комп'ютерні науки .– 2012. – Том 138. – С. 35–38.

11. Transforming Society through Technological Innovation. Режим доступу до сайту: http://www.cmu.edu/homepage/computing/2010/fall/nell-computer-that-learns.shtml.

12. Марченко А.А. Контекстный семантический анализ текста. Система текстового мониторинга и качественного оценивания фокусного объекта / А.А. Марченко, А. А. Никоненко// Искусственный интеллект. − 2008. − Вип.3.− С. 808−813.

13. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval, Vol. 2, No 1-2 (2008) p.1–135.

Рецензент: д.ф.-м.н., Терещенко В.М. професор кафедри математичної інформатики факультету кібернетики КНУ імені Тараса Шевченка.

д.ф.-м.н., доц. Марченко О.О., к.т.н. Марченко-Бабич О.Н.

РАЗРАБОТКА ЛИНГВИСТИЧЕСКОГО ОБЕСПЕЧЕНИЯ ДЛЯ АВТОМАТИЗИРОВАННОГО МОНИТОРИНГА СОЦИАЛЬНЫХ СЕТЕЙ

В статье рассматриваются новые принципы и подходы к разработке лингвистического обеспечения для автоматизированного семантического поиска и отбора информации, содержание которой может свидетельствовать о целенаправленном воздействии на аудиторию, среди текстов социальных сетей и других социально-ориентированных Интернет-сервисов. Установлены особенности языковых конструкций и компонентов базы знаний системы автоматизированного мониторинга. Предложен способ повышения точности смыслового анализа текстов, в основу которого входит уход от влияния полисемии многозначных слов и словосочетаний. Подходы, изложенные в статье, дают возможность получить в процессе поиска текст, который не содержит ни одного ключевого слова из запроса и при этом полностью отвечает ему по смыслу и семантике. Предложены виды анализа текста, которые целесообразно использовать для семантического поиска в текстах социальных сетей и других социально-ориентированных Интернет-сервисах: контекстный ассоциативно-семантический и Sentiment Analysis, что дает возможность учета многих деталей и нюансов эмоциональной окраски текстовых сообщений и может отображать общую направленность информационного воздействия, которое осуществляется заинтересованными участниками информационного пространства.

Ключевые слова: информационное воздействие, информационное пространство, социальная сеть, Интернет-сервис, обработка естественного языка, алгоритм поиска.

As.Prof. Marchenko O.O., Ph.D. Marchenko-Babic O.N.

DEVELOPING PROVIDING FOR LINGUISTIC THE AUTOMATED MONITORING OF SOCIAL NETWORKS

The article examines new principles and approaches to linguistic support for automatized semantic retrieval and information sampling among social networks and other social-oriented Internet-services. The information contents can signify of directed influence towards the audience. The features of linguistic structures and knowledgebase components of the automatized monitoring system are defined. The technique of increase of texts semantic analysis precision is suggested. It comprise avoidance of the impact of the multivalued wards and phrases polysemy. The article approaches let obtain the text without any keyword of the query but in the meantime completely corresponding to it in the process of the retrieval. Including of the context association-semantic analysis and Sentiment Analysis is expedient in the process of semantic retrieval of the social network and other social-oriented Internet-services. These kinds of analysis let regard many details of the text messages emotional coloration and are able to reflect the total tendency of the information influence from the interested parts of the information environment.

Keywords: information influence, information environment, social network, Internet-service, natural language processing, retrieval algorithm.