Правова інформатика

УДК 004.912

ЛАНДЕ Д.В., *доктор технічних наук, професор*, керівник Наукового центру інформатики і права ДНУ ІІБП НАПрН України,
завідувач кафедри НН ФТІ КПІ ім. Ігоря Сікорського.
ORCID: https://orcid.org/0000-0003-3945-1178.
ШНУРКО-ТАБАКОВА Е.В., голова ГО Ради інформбезпеки та кіберзахисту.

МЕРЕЖЕВИЙ АНАЛІЗ СУСПІЛЬНОЇ ДУМКИ ЩОДО ПРАВ ЛЮДИНИ В КРАЇНІ-АГРЕСОРІ

Анотація. *Наведено методику автоматичного виділення та ідентифікації зв'язків фразеологізмів в інформаційних потоках з метою подальшої ідентифікації наративів як узагальнення набору фразеологічних одиниць. За допомогою сучасних методів мережевого аналізу досліджуються зв'язки фразеологізмів та виділяються їх окремі кластери, які ймовірно відповідають наративам. Також запропоновано форму візуального відображення інформаційного потоку фразеологізмами та датами. Як приклад наведено застосування методики для аналізу уявлень росіян, які відображаються в соціальних мережах, щодо прав людини.*

Ключові слова: *інформаційні потоки, права людини, сталі словосполучення, виявлення понять, мережа понять, візуалізація.*

Summary. *The technique of automatic allocation and identification of connections of phraseological units in information streams for the purpose of further identification of narratives as a generalization of a set of phraseological units is given. With the help of modern methods of network analysis, the connections of phraseological units are studied and their separate clusters are determined, which probably correspond to narratives. A form of visual display of the information flow by phraseological units and dates is also proposed. As an example, the application of the methodology for the analysis of the views of russians regarding human rights, which are reflected in social networks, is given.*

Keywords: *information streams, human rights, set phrases, concept extraction, concept network, visualization.*

Постановка проблеми. Для проведення аналітичних досліджень на базі застосування соціальних мереж вивчаються документи, повідомлення, які надходять туди від учасників, користувачів. У питанні дослідження публікацій в соціальних мережах має значення не лише відстеження відомих тем/публікацій, але й є актуальним питанням всебічного дослідження контенту та виявлення актуальних тем, наративів та сюжетів. Аналіз великих масивів публікацій визначених сегментів Інтернету (сайти, соціальні мережі) дозволяє, зокрема, виявляти непомічені наративи на стадіях їх розвитку, а також забезпечувати інформаційну та технічну протидію агресору на всіх стадіях розгортання інформаційних операцій, кампаній [1]. Аналіз динаміки наративів дозволяє прогнозувати суспільні явища, події та процеси. Дослідження штучно створених і впроваджених наративів може виявити ймовірні приховані цілі ворога. Дослідження старих наративів дозволяє перейти від ретроспективного аналізу до прогнозу. Виявлення нових наративів може полегшити та сприяти оперативному розвідуванню, виявленню ворожих намірів. Тому завдання дослідження наративів є особливо актуальним у контексті інформаційних та гібридних війн.

Метою статті є представлення методики мережевого аналізу суспільної думки в правовій сфері.

Виклад основного матеріалу. Суть методики полягає у виконанні експертами такої технологічної операції, як створення запиту, що відповідає об'єкту, що цікавить, до існуючих інформаційно-пошукових систем [2; 3]. В результаті обробки цих запитів створюються великі набори релевантних документів, в яких за допомогою спеціальних алгоритмів ідентифікуються необхідні фрагменти. На основі відібраних для різних мов корпусів виділено постійні словосполучення, що належать до різних часових періодів. Інструменти та засоби ідентифікації та виділення стійких словосполучень базуються на концепціях машинного навчання, лінгвістичного аналізу та статистичних розрахунків. Далі за допомогою сучасних методів графового аналізу досліджуються зв'язки фразеологізмів та виділяються їх окремі кластери , які ймовірно відповідають наративам.

Запропонована в даній роботі методика виділення, дослідження динаміки та виявлення зв'язків фразеологізмів в інформаційних потоках передбачає реалізацію ряду етапів, а саме:

Крок 1. Формування початкових запитів (шаблонів для виділення тексту), що відповідають загальній темі.

Для проведення досліджень було використано інформацію із соціальних мереж за певний період. Масив даних має бути попередньо сформований у результаті технологічних операцій пошуку/відбору документів за експертними запитами, що змістовно відповідають тематиці дослідження. В наведеному прикладі досліджуються уявлення російських користувачів щодо прав людини у розрізі проблематики війни, що розв'язана РФ проти України. Зокрема, для аналізу застосовувався запит до сервісу моніторингу соціальних медіа із визначенням необхідного діапазона дат:

((наруш~/1/прав-человек)|(защи~/1/прав-человек)|правозащ) украин

У результаті опрацювання запиту отримується масив документів із інформаційних джерел – соціальних медіа (Рис. 1).

№	Джерело	Повідомлень
1	Youtube	4364
2	Telegram	1772
3	ВКонтакте	365
4	LiveInternet.ru	104
5	Одноклассники	59
6	LiveJournal	53
7	RuTube	25
8	Facebook	18

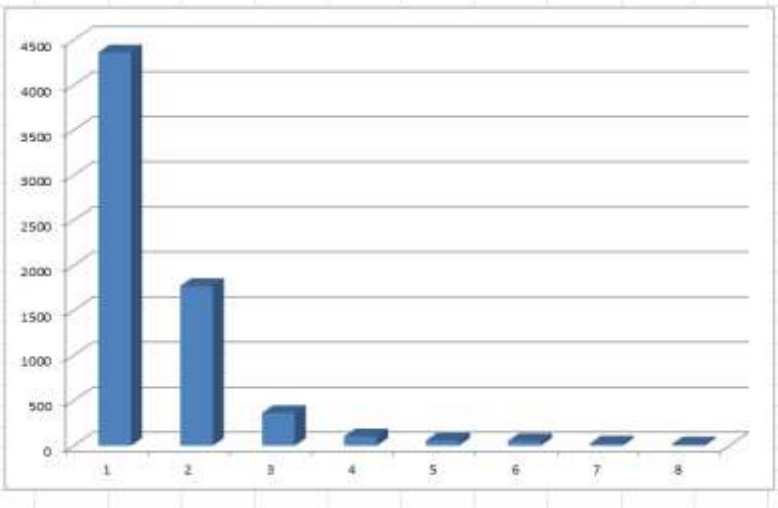

Рис. 1 – Кількість тематичних повідомлень і гістограма розподілу кількості повідомлень в соціальних мережах за вибраний період (01.02.2021 – 07.08.2022).

У результаті опрацювання запиту отримується інформаційний масив обсягом у 7050 повідомлень, щодобову динаміку кількості публікацій в якому наведено на Рис. 2.

Рис. 2 – Статистика повідомлень за запитом.

Після цього досліджуються піки на цій діаграмі, визначаються події, що відповідають цим пікам, зокрема, у наведеному прикладі (заголовки мовою оригіналу):

2022.04.07: Генасамблея ООН призупинила участь Росії в Раді із прав людини.

2022.08.04: Теракт РФ в Еленовке.

Крок 2. У результаті обробки запитів, створених на першому кроці, створюються великі набори релевантних документів, у яких за допомогою спеціальних алгоритмів необхідно визначити необхідні фрагменти. У цьому випадку фрагментами є речення, які відповідають запиту. Крім того, для кожного такого речення до вихідного файлу також додаються сусідні речення.

Крок 3. Суть третього кроку полягає у вилученні найбільш важливих окремих слів (уніграм) і фраз з файлів, отриманих на кроці 2. Для цього запускається програмний комплекс для вилучення ключових термінів (слів і словосполучення) із тематичних інформаційних потоків для подальшої ідентифікації фразеологізмів [4]. Цей модуль призначений для попередньої обробки текстових даних природною мовою тематичних інформаційних потоків. Попередня обробка включає токенізацію тексту, видалення стоп-слів і подальше виділення ключових слів і ключових фраз шляхом застосування спеціальних методів обробки природної мови. Таким чином, на базі отриманого інформаційного масиву визначаються всі сталі словосполучення – базові словоформи. У нашому прикладі це найбільш вживані словосполучення, що відповідають правам людини, які порушуються з погляду росіян (мовою, якою наведені повідомлення):

право на жизнь
право на самоопределение
првао на виртуальные активы
право на свободу перемещения | право на свободное передвижение
право на восстание
право на аборт
право на самооборону | право на индивидуальную или коллективную самооборону
право на свободу собраний | право на свободу мирных собраний
право на свободу слова | право на свободу мысли
право на юридическую помощь
право на труд
право на расторжение контракта

право на проведение митингов
право на образование
право на льготный проезд
право на свободу вероисповедания
право на образование на родном языке
право на эксплуатацию естественных ресурсов
право на территориальную целостность
право на свободу собраний
право на справедливый суд
право на службу без оружия
право на гуманное обращение
право на защиту и справедливый суд
право на владение
право на свободный въезд и выезд
право на применение силы
право на мирную жизнь
право на медицинскую помощь
право на безвизовый въезд
право на бесплатный транзит
право на свободу выражения мнений
право на свободу и личную неприкосновенность
право на получение информации

Крок 4. Далі для кожного із видів фразеологізмів (сталих словоформ) шляхом уточнення первинного запиту отримуються часткові інформаційні масиви релевантних документів, із яких автоматично екстрагуються прізвища персон, що асоціюються авторами повідомлень із відповідними правами (або, найчастіше, з їх порушеннями).

Крок 5. На цьому кроці застосовується форма візуального відображення інформаційного масиву у розрізі фразеологізмів і дат. Цю форму представлено діаграмою – прямокутною таблицею, клітинки якої заповнюються числовими значеннями, що відповідають частотності фразеологізмів по відношенню до дат, коли вони з'являються [5]. Тобто стовпці цієї таблиці відповідають датам, а рядки – фразеологізмам, які можна розглядати як своєрідні змістовні фільтри інформаційного масиву. Передбачається, що комірки зафарбовані відтінками кольорів, залежно від значень обсягів публікацій за вибраним фразеологізмом у відповідний день. Запропоновані діаграми дозволяють без додаткової обробки виявляти групи найбільш пов'язаних за датами та інтенсивністю публікацій об'єктів візуально.

Для подальшого групування (кластеризації) формується мережа взаємозв'язків видів прав і персон, в якій визначаються групи найбільш зв'язаних між собою і віддалених від інших. Передбачається, що отримані кластери із тісно пов'язаних фразеологізмів і відповідатимуть наративам. На практиці форма реалізована у вигляді html-файлу із застосуванням мови JavaScript. Приклад такого відображення наведено на Рис. 3. На цій діаграмі яскраві горизонтальні лінії (висока частотність окремих фразеологізмів протягом певного періоду) можуть чітко підказати користувачеві про тенденції громадської думки.

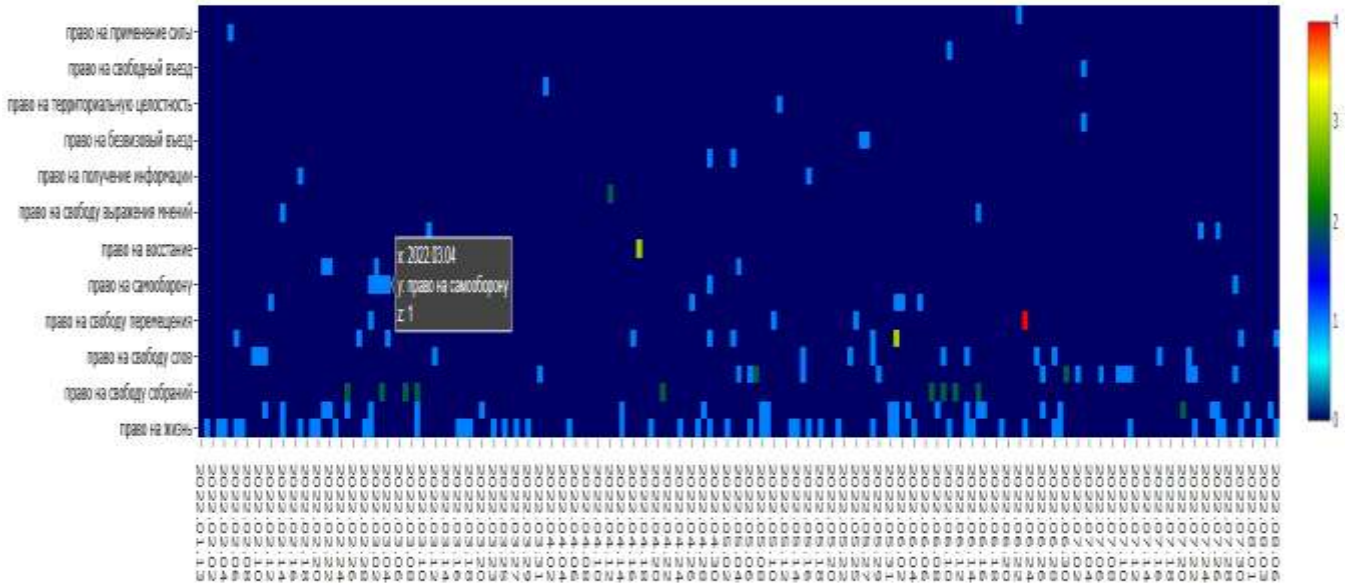

Рис. 3 – Діаграма відображення фразеологізмів за часом.

Крок 6. Наступним кроком за допомогою методів теорії графів досліджуються взаємозв'язки персон і фразеологізмів, що відповідають видам прав людини. Окремі поняття вважаються пов'язаними, якщо вони одночасно є частиною одного тексту. Вагомість таких відношень між двома поняттями відповідає кількості фрагментів тексту документів, які відповідають двом пов'язаним фразеологізмам одночасно. На основі сформованої матриці виводиться відповідний граф, інакше кажучи, семантична мережа.

Крок 7. На останньому кроці мережа кластеризується [6]. Кластеризація може здійснюватися за допомогою різних алгоритмів. Зокрема, можуть бути використані алгоритми, які базуються на розрахунку модульності мережі [7]. На основі кластеризації, отриманої за допомогою експертного аналізу, можна ідентифікувати наративи, які є узагальненнями фразеологізмів, які разом є частиною тих самих кластерів. У нашому прикладі урахування контексту появи окремих видів прав людини і прізвищ персон побудовано мережу, яку кластеризовано (виділено найбільш зв'язні групи понять). Всього визначено 5 основних кластерів.

Приклад візуалізації мережі взаємозв'язків понять за допомогою програми Gephi демонструє кластери на основі всього інформаційного масиву див. Рис. 4.

Кластери виділені окремими кольорами.

Зміст окремих кластерів визначається експертом. Таким чином, в результаті автоматизованого аналізу отримуються основні кластери побудованої мережі, найвагоміші поняття серед яких:

1. Право на самовизначення.
2. Право на проведення мітингів
3. Право на свободу і особисту недоторканність
4. Право на свободу зборів
5. Право на свободу слова

Після цього визначаються центральні поняття отриманих кластерів, на основі яких формуються ланцюжки сюжетів, що їм відповідають.

Ранжируваний перелік цих ланцюжків є основою подальшого формування інформаційних звітів, дайджестів.

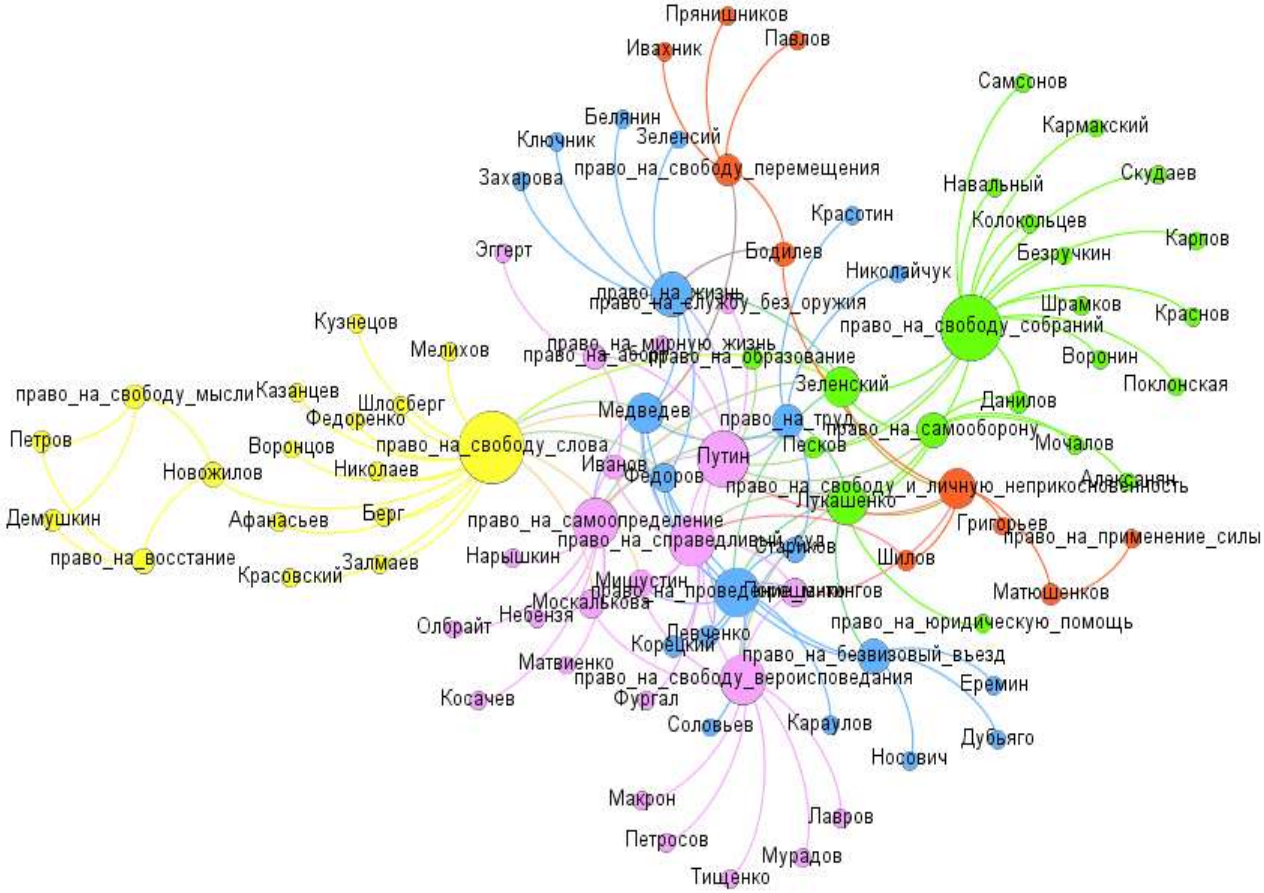

Рис. 4 – Приклад візуалізації мережі взаємозв’язків понять за допомогою програми Gephi демонструє кластери на основі всього інформаційного масиву.

Висновки.

Запропонований лінгвостатистичний підхід проведення аналітичних досліджень на базі застосування соціальних мереж має об’єктивний характер, що є суттєвим компонентом методологічної основи аналізу та прогнозування.

Пропонована методика не вимагає великих людських і часових ресурсів. Адже основні витрати вже були внесені користувачами соціальних мереж.

У результаті експериментів є підстави припустити, що використання запропонованих засобів візуалізації дозволяє “розкласти” вихідні часові ряди за складом і особливостями фразеології і понять, виявити активність публікацій, яким відповідають певні наративи, виявити зв’язки фразеологізмів, особливості динаміки виникнення в інформаційному потоці нових фразеологізмів.

Розглянутий підхід може бути використаний для аналізу та візуалізації розподілу наративів для будь-яких вибраних наборів інформації з точки зору питань, що цікавлять дослідника та охоплюють значний часовий проміжок.

Запропонована методологія знайшла практичне втілення в реальних дослідженнях з проблематики прав людини, що проводяться аналітичною групою Index Systems, зокрема тих, які спрямовані на виявлення реального стану справ зовнішньої та внутрішньої політики РФ як країни-агресора, на базі сервісу моніторингу та автоматизованого визначення інформаційних загроз attackindex.com.

Використана література

1. Горбулін В.П., Додонов О.Г., Ланде Д.В. Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія. Київ: Інтертехнологія, 2009. 164 с. ISBN 978-966-164-81-27.

2. Manning C.D., Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press, 2018. 482 p. ISBN 978-052-186-57-15, 052-18-657-19.

3. Згуровський М. та ін. (2022). Підвищення актуальності пошуку інформації в Інтернет-медіа та соціальних мережах у завданнях планування сценаріїв. In: Zgurovsky, M., Pankratova, N. (eds) System Analysis & Intelligent Computing. SAIC 2020. Дослідження обчислювального інтелекту, том 1022. Springer, Cham. DOI: 10.1007/978-3-030-94910-5_10

4. Ланде Д., Дмитренко О. Використання тегування частин мови для побудови мереж термінів у правовій сфері: матеріали 5-ї Міжнародної конференції з комп'ютерної лінгвістики та інтелектуальних систем (COLINS 2021). Т. I: Основна конференція Львів, 22-23 квіт. 2021 р. Матеріали семінару CEUR (ceur-ws.org). Т. 2870. С. 87-97. ISSN 1613-0073.

5. Zgurovsky M., Lande D., Yefremov K., Dmytrenko O., Boldak A., Soboliev A. Extracting and Identifying Relationships of Key Phrases in Information Flows. Published in: 2022 IEEE 3rd International Conference on System Analysis & Intelligent Computing (SAIC) 04-07 October 2022. DOI: 10.11 09/SAIC57818.2022.9923019

6. Ланде Д., Субач І., Пучков О., Соболєв А. Метод кластеризації для узагальнення інформації та моделювання інформації та безпеки предметної області. Міжнародний журнал 50, № 1 (2021): 79-86. DOI: 10.11610/isij.5013.

7. Cherven K. Mastering Gephi Network Visualization. Packt Publishing, 2015. 378 p. ISBN 781-783-98-73-44.

~~~~~~~~~~~~~ * * * ~~~~~~~~~~~~~
~~~~~~~~~~~~~