УДК 316

Кириченко, Р. (2021). Типологія задач машинного аналізу текстів у сучасній соціології. *Соціологічні студії*, 2 (19), 53–62. https://doi.org/10.29038/2306-3971-2021-02-41-48

Типологія задач машинного аналізу текстів у сучасній соціології

Роман Кириченко –
здобувач PhD, ОП «Соціологія», факультет соціології, Київський національний університет імені Тараса Шевченка,Україна

Roman Kyrychenko –
student of PhD program, Faculty of Sociology, Taras Shevchenko National University of Kyiv, Ukraine
E-mail: kirichenko17roman@gmail.com,
ORCID: https://orcid.org/0000-0002-1610-4352

DOI: 10.29038/2306-3971-2021-02-41-48

Received: September, 2021
1st Revision: October, 2021
Accepted: October, 2021

У статті розглянуто можливості використання сучасних методів обробки текстів для соціологічного аналізу. Основну увагу приділено трьом завданням, які наразі можна виконати засобами обчислювального аналізу текстів: аналіз змістовної близькості, моделювання тем та сентимент-аналіз (аналіз тональностей). В останні роки методи обробки природної мови настільки прогресували, що це дає змогу соціологам автоматично фіксувати семантику текстів, порівнювати її в часі, групувати на підставі схожості. Також це уможливлює масштабування аналізу великих масивів документів, що відкриває нову сторінку в розвитку контент-аналізу, за якої ми наближаємося до відмови від ручного кодування документів, а дослідники зможуть сконцентруватися на аналізі. Ми продемонстрували ці можливості на прикладі аналізу новин із ресурсу «Українська правда» за 2001–2020 рр. Методи, застосовані в статті, дали нам змогу повністю автоматизовано виявити, які семантичні зрушення щодо слів, пов'язаних із діяльністю правоохоронних органів, відбувалися під дією соціальних факторів протягом останніх двадцяти років. Також ми згрупували новини за основними темами повідомлень про поліцію в матеріалах видання й проаналізували, чи змінювалося ставлення до неї протягом його існування.

Ключові слова: обчислювальний аналіз текстів, аналіз змістовної близькості, моделювання тем, сентимент-аналіз.

Kyrychenko Roman. Typology of Tasks of Machine Analysis of Texts in Contemporary Sociology. This article considers the possibilities of using modern methods of word processing for sociological analysis. The main focus is on three tasks that we can currently solve using computational analysis of texts: analysis of semantic proximity, modeling of themes, and sentiment analysis. The methods discussed in this article have helped us to fully automate the semantic shifts in law enforcement-related words over the past twenty years. In recent years, the methods of processing natural language have progressed so much that it allows sociologists to automatically record the semantics of texts, compare them over time, and group based on similarity. It also allows us to scale the analysis of large arrays of documents, which opens a new page in the development of content analysis, in which we are approaching the abandonment of manual coding of documents, and researchers will be able to focus on study. We demonstrated these capabilities based on the news analysis from the resource «Ukrainska Pravda» for 2001–2020. We also grouped the news on the main topics of police reports in the publication materials and analyzed whether attitudes towards it changed during its existence.

Key words: computational analysis of texts, content proximity analysis, topic modeling, sentiment analysis.

ВСТУП

Дослідження документів у соціології від початку її існування є популярним методом, адже тексти є відображенням соціальної дійсності (Lemke, Wiedemann, 2016). Однак метод спирається переважно на ручне кодування дослідниками або досить примітивний аналіз частот. До 1950-х аналіз текстів був переважно описовим і спирався на частотний аналіз. Зараз усе більше інформації генерується в мережі Інтернет, що робить її більш доступною для дослідників. Проте великі обсяги документів важче кодувати. Отже, виникає суперечність між обсягом документів, доступним дослідникам, який дуже швидко зростає, та реальними можливостями вчених обробляти такі дані. Відтак актуалізуються дослідження автоматизації вивчення документів у соціології.

У цьому нам можуть допомогти методи обробки природної мови, запозичені з машинного навчання. Останнім часом ця сфера засвідчує зростання й існує потреба в систематизації розуміння того, як можна використати ці методи для соціологічного аналізу документів, зокрема для автоматизованого контент-аналізу. Із цього й очевидна **мета** цієї статті – показати, які завдання аналізу документів можна виконувати за допомогою новітніх методів обробки природної мови. При цьому основний акцент здійснюватиметься на методи, які пропонують максимальне зменшення втручання дослідника в їхню роботу (насамперед це непараметричні методи).

Наукова новизна матеріалів, викладених у статті, для української соціології безсумнівна, оскільки ці методи вперше застосовуються для соціологічного аналізу документів, написаних українською мовою.

1. КОРОТКА ІСТОРІЯ МЕТОДІВ АВТОМАТИЗОВАНОГО АНАЛІЗУ ПРИРОДНОЇ МОВИ

Практично від часу появи програмованого комп'ютера дослідники намагалися його використати для аналізу мови. Першим великим важливим проєктом у цьому напрямі був *General Inquirer* (розробка Гарварду) – програма, що займалася класифікацією й частотним аналізом слів та фраз за категоріями (Stone et al., 1966). Загалом там представлено 182 категорії[1], які діляться на 26 блоків. Ці категорії створені на основі словників і фактично є набором слів, що належать до певних тем/сентиментів/цінностей.

Комп'ютерна обробка текстів у соціології пройшла декілька етапів еволюції. Обробляти тексти комп'ютером для дослідницьких цілей почали ще в 1960-х роках. Ця обробка тривалий час обмежувалася просто фіксацією наявності певних слів у текстах та підрахунком їх частоти (від цього й зараз не відмовляються). Із появою глобальної мережі Інтернет і зростанням кількості текстових масивів з'являються складніші методи обробки мови. Насамперед це методи мішка слів (*BoW – bag of words*[2]), поєднані з Баєсовськими моделями.

Донедавна такі методи давали змогу непогано кодувати символічну сторону текстів, однак не завжди правильно відображали зміст. Насамперед гострою була проблема врахування контексту слова. Частково це намагалися розв'язати, застосовуючи біграм/триграм-моделі, але вони враховували лише локальний контекст, а для складних документів цього замало.

Однак в останні п'ять років ситуація кардинально змінилася й з'явилися моделі, спроможні гарно враховувати контекст. Передусім це пов'язано з моделями векторного представлення слів (*word embeddings*), які кардинально змінили підхід до кодування текстів[3].

Самі підходи до кодування прекрасно ілюструють те, як еволюціонували методи автоматизованої обробки текстів. Зрозуміло, що якщо ми хочемо здійснити будь-яке кількісне дослідження текстів, то нам потрібно певним чином конвертувати їх у числове представлення. Найпростіший спосіб – зробити загальний словник відомих слів і для кожного слова в кожному документі порахувати частоту його появи. Такий підхід більш-менш якісно дає нам змогу уявити те, про що йдеться в тексті, однак, наприклад, здійснити аналіз сентиментів уже складніше, оскільки для нього потрібно враховувати порядок слів, зокрема наявність заперечення. Цей недолік можна частково виправити, використовуючи біграми/триграми, тобто рахувати не наявність певних слів, а комбінацій із двох/трьох слів. Цей підхід лише частково розв'язував проблему й робив це екстенсивним шляхом (закодований масив текстів через такий спосіб кодування дуже зростав, а отримана таблиця була надто розрідженою). Також не вирішено питання загальної популярності окремих слів. Деякі слова (сполучники, займенники) трапляються практично в кожному тексті й мало сприяють розумінню особливостей текстів. Аби розв'язати цю проблему, придумали рахувати

[1] http://www.wjh.harvard.edu/~inquirer/

[2] Такий спосіб репрезентації текстів, де зберігається лише інформація про кількість появи кожного слова в тексті, натомість ігноруються дані про їх порядок (Harris, 1954).

[3] Наприклад, це представлення може мати такий вигляд:

dim 1	dim 2	dim n
-1.500	2.230	-1.010
3.100	-0.170	2.540
0.170	0.001	-2.980

частоту кожного слова інакше – через метрику *TF-IDF* (term frequency – inverse document frequency[1]), що є результатом ділення частоти появи слова в конкретному документі на число документів у корпусі текстів, де це слово присутнє. Такий підхід дав змогу показати, що певні слова виділяють один текст із-поміж інших. Тому що якщо одне слово наявне у великій кількості лише в декількох документах, то його значення для них буде дуже високим. Натомість якби воно траплялося так само часто в більшості документів, то його значення було б нівельоване. Ці підходи дуже прості й легкозрозумілі, однак вони досить поверхово передають змістове навантаження тексту.

Текст – це складна система, яка часто містить у собі латентні змінні, котрі важко виразити просто через частоту слів чи їх комбінації. Потрібно досить точно оцінювати змістову близькість слів і їх поєднань. І такі методи з'явились у 2013–2014 рр., коли опубліковано моделі *word2vec* (Mikolov et al., 2013) та *GloVe* (Pennington et al., 2014), які кардинально змінили підхід до кодування слів у текстах. Ця кардинальна зміна полягає в тому, що повністю відмовилися від кодування частот появи слів, натомість вони тепер використовуються як попередній етап до кодування. Ідея цих двох моделей полягає в тому, щоб, замість частот слів, текст характеризувався набором векторів, які характеризують зміст цих слів. Тому спочатку на масиві всіх текстів аналізуються сусідства слів (зазвичай, близькі за сусідством слова є частіше близькими за змістом) і будуються моделі, які 1) або на основі контексту намагаються вгадати слово по сусідству (*CBOW*); 2) або на основі слова намагаються відтворити його контекст (набір сусідніх слів) (*skip-gram*). Матриці ваг і систематичних помилок нейронної мережі, котра вивчає такі моделі, далі використовують як таблицю з координатами слів у n-мірному векторному просторі.

Перевагою такої репрезентації слів у текстах є можливість робити широкий аспект математичних операцій. Наприклад, охарактеризувати зміст речення також можна через n-мірний вектор, просто використавши середні значення кожної координати кожного слова в реченні. Для соціологів головний наслідок, що, незалежно від написання, слова можуть мати як дуже близькі векторні значення, так і дуже далекі (наприклад, слова «соціологія» і «Вебер», попри 0 однакових букв у слові, така модель постає ближче в просторі, ніж «соціологія» й «соціоніка»). Звісно, таке кодування теж має недоліки (наприклад, воно дуже витратне в часі та все одно не справляється з омонімами), однак у подальші роки над ним продовжувалась активна робота та з'явилося безліч варіантів стосовно того, як ці недоліки можна виправити. Для нас більш важливий вибух методів обробки природної мови, які такий спосіб кодування використали. Спочатку це були моделі рекурентних нейронних мереж *RRN, LSTM, GRU*[2] (Yin et al., 2017), а з 2018 року в цій сфері запанували моделі-трансформери, найвідомішими представниками яких є моделі *BERT* (Devlin et al., 2019) та *GPT2* (Radford et al.)[3].

Ще більше значення має те, що кодування слів/документів через векторне представлення, по суті, характеризує смислові одиниці тексту через латентні ознаки, які, зі свого боку, цікавлять дослідників (цими латентними ознаками можуть бути тематики, сентименти текстів).

Обчислювальний аналіз текстів у соціології ще не став загальнопоширеним і стандартизованим. Значною мірою це спричинено тим, що нові підходи до аналізу з'явилися лише в останні роки. Однак деякі дослідники, такі як Мейрінг, Лемке та Відеман, звертають увагу на те, що аналіз природної мови в соціології на сьогодні не є стандартизованим і систематизованим (Lemke & Wiedemann, 2016). На їхню думку, існує необхідність у встановленні універсальних стандартів обчислювального аналізу текстів для того, щоб дослідники більше фокусувалися на актуальних дослідженнях, а не на розробці нових методів (Lemke & Wiedemann, 2016).

---

[1] tf-idf ваги складаються з двох компонентів: 1) Term Frequency (TF) – кількість потраплянь слова в документ / кількість слів у документі $\mathrm{tf}(t,d) = \frac{n_t}{\sum_k n_k}$; 2) Inverse Document Frequency (IDF) – логарифм кількості документів у корпусі / кількість документів, що містять слово $\mathrm{idf}(t,D) = \log \frac{|D|}{|\{d_i \in D | t \in d_i\}|}$. Відповідно, tf-idf рахується як $\text{tf-idf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$.

[2] Ці архітектури нейронних мереж дали змогу якісніше обробляти послідовності текстів, оскільки для них характерно те, що вивід моделі є одночасно і входом на наступну ітерацію її роботи.

[3] Головною новацією цих архітектур нейронних мереж є механізм уваги (attention), який дає змогу встановити зв'язки між одиницями в тексті навіть у тих випадках, коли близько пов'язані слова не стоять поруч (Vaswani et al., 2017).

Отже, наразі існує потреба в систематизації завдань та методів обробки природної мови для соціологічного аналізу документів.

2. ЗАВДАННЯ ОБРОБКИ ПРИРОДНОЇ МОВИ ТА ЇХ ВИКОРИСТАННЯ В СОЦІОЛОГІЧНОМУ АНАЛІЗІ ДОКУМЕНТІВ

Аналіз досліджень із використанням методів обробки природної мови, результати яких опубліковані в останні роки, дав підставу виділити три основні класи завдань аналізу, які виконують соціологи засобами обробки природної мови:

- аналіз зміни сенсів слів;
- моделювання тем;
- семантичний аналіз.

Рис. 1. *Завдання аналізу документів*
Джерело: автор.

2.1. Аналіз зміни сенсів слів

Мова – мінливий соціальний продукт. Значення одних і тих самих слів може змінюватися в часі. Наприклад, англійське слово «apple» зараз переважно асоціюється з однойменною компанією, хоча раніше це слово вживалося лише для позначення фрукта (Yao et al., 2018).

Векторне представлення слів спроможне кількісно продемонструвати, як значення слів змінюється. Це пов'язано з тим, що воно представляє слова таким чином, що ми можемо з'ясувати їх положення в змістовому просторі відносно інших слів. Наприклад, у випадку з вищезгаданим словом «apple» векторне представлення побудоване на текстах XIX ст. показувало його близькість до слів, пов'язаних із фруктами, деревами, рослинами. Тоді як векторне представлення, зроблене на текстах сучасності, розмістило в просторі це слово поруч зі словами, пов'язаними з комп'ютерами, інформаційними технологіями, корпораціями. Ось ці зміщення в просторі значень є цікавими для аналізу, адже вони ілюструють, як історично змінювалося сприйняття певного концепту. Так само можна досліджувати, як варіюється значення поняття залежно від групового контексту.

На сьогодні існує декілька технік, які застосовують векторне представлення слів для аналізу еволюції значень[1]. Найпростішим способом було б просто вивчити окремі векторні представлення для кожного часового періоду й порівнювати положення кожного слова. Але проблема в тому, що в кожному корпусі наявні відмінності в словниках (не всі слова трапляються в різні періоди) і векторне представлення розраховується суто відносно інших слів, тому тут мають значення не конкретні координати слова в n-мірному просторі, а їх віддаленість від координат інших слів (простіше кажучи, системи координат у різних векторних представлень можуть суттєво відрізнятися). Тому існує потреба в приведенні окремих векторних представлень слів до однієї системи координат так, аби їх можна було коректно порівнювати.

Бамлер і Мандт у статті «Динамічні векторні представлення слів» аналізують еволюцію значення слів протягом 150 років з архіву книг Google за допомогою власної версії динамічного

[1] У роботі через різні варіанти векторного представлення слів (word2vec na GloVe) проаналізовано геометрію класів в американському суспільстві. Досліджено, які концепти асоціюються з різними соціальними класами й гендерами (Kozlowski et al., 2018).

векторного представлення слів, яка є розширенням оригінальної скіп-грам моделі Міколова. Це розширення полягає в додаванні в модель латентного часового ряду, який дає змогу, з одного боку, вивчити модель, яка б ураховувала зміну векторних значень слів у просторі, а з іншого – реалізовувала це через навчання всього одного векторного представлення, замість набору моделей для різних періодів часу (Bamler & Mandt, 2017).

Джанг у дисертації «Динамічні векторні представлення слів для аналізу новин» інакше розв'язав проблему динамічності векторних представлень. Його алгоритм полягав у навчанні окремих векторних представлень для кожного часового періоду й подальшому їх обертанні для мінімізації відмінностей у координатах одних і тих самих слів. Цей підхід використано для аналізу динаміки слів «Трамп» та «Хорватія» у ЗМІ протягом 2018 р. (Zhang, 2019).

Яо, Сун, Дін, Рао і Сюн у статті «Динамічні векторні представлення слів для еволюціонуючих семантичних досліджень» запропонували свій варіант динамічного векторного представлення слів, який теж ґрунтувався на обертанні просторів для накладання різних векторних представлень у часі. Проаналізовано близько 100 тисяч статей із «Нью-Йорк Таймс» за період із 1990 по 2016 р. Їх основним фокусом уваги було знаходження еквівалентних нині слів у минулому часі. Наприклад, виявлено, що в 1990-х роках місце слова «твіттер» займали слова «телебачення», «радіо», «повідомлення», тобто були його функціональними еквівалентами в цей час (Yao et al., 2018).

Ді Карло, Б'янкі та Пальмонарі в статті «Навчання часових векторних представлень слів з компасом» представили альтернативний метод навчання векторного представлення слів, яке б могло показувати семантичну динаміку. Головна ідея такого підходу полягає в навчанні векторного представлення слів за два етапи. На першому досліджується загальне векторне представлення для текстів за всі роки, які планується розглянути. Це треба для того, щоб створити простір, відносно якого вимірюватимуться зміни значень за роками (автори називають це векторне представлення «компасом»). Потім уже вивчаються векторні представлення за часові періоди, які дослідники планують порівнювати. Однак значення в просторі для слів розраховують відносно їх же значень у «компасі». Цього ефекту вдається досягти завдяки тому, що кожне векторне представлення слів для часового періоду вчиться не із самого початку, а від своїх значень у векторному представленні-компасі (ці значення ініціалізуються перед навчанням кожного нового векторного представлення, а це пришвидшує процес навчання моделі). Усі вищезгадані моделі розглянуто на основі архітектури word2vec (Di Carlo et al., 2019).

Використано підхід TWEC для навчання векторних представлень слів за окремі роки на основі новинних повідомлень з інтернет-видання «Українська правда». Ми взяли новини за 2001–2020 рр. й установили окреме векторне представлення для кожного року (загалом проаналізовано 314 436 новинних повідомлень).

Таблиця 1

Кількість аналізованих новин за роками

Рік	Кількість	Рік	Кількість
2001	4497	2011	17 171
2002	6085	2012	16 599
2003	4969	2013	15 898
2004	8395	2014	24 045
2005	9923	2015	25 790
2006	11 748	2016	23 648
2007	13 995	2017	21 286
2008	14 930	2018	20 607
2009	15 627	2019	19 093
2010	16 538	2020	23 592

Джерело: авторські дані.

Завдяки цим моделям ми можемо побачити, наприклад, як еволюціонувало значення слова «поліція». В останні роки найбільш семантично близькими до цього поняття є слова «нацполіція» та «прокуратура». Також посилились ототожнення поліції з владою/адміністрацією. Відходить у минуле ототожнення поліції й міліції. Цікаво, що у 2013–2015 рр. поліція також сильно асоціювалася зі словом «самооборона», що, вочевидь, відображає діяльність загонів народної самооборони в часи

Євромайдану. Також відходить у минуле стійка асоціація поліції з міліцією, що, найімовірніше, є наслідком реформи цієї структури та офіційним перетворенням міліції в поліцію згідно із законом «Про національну поліцію».

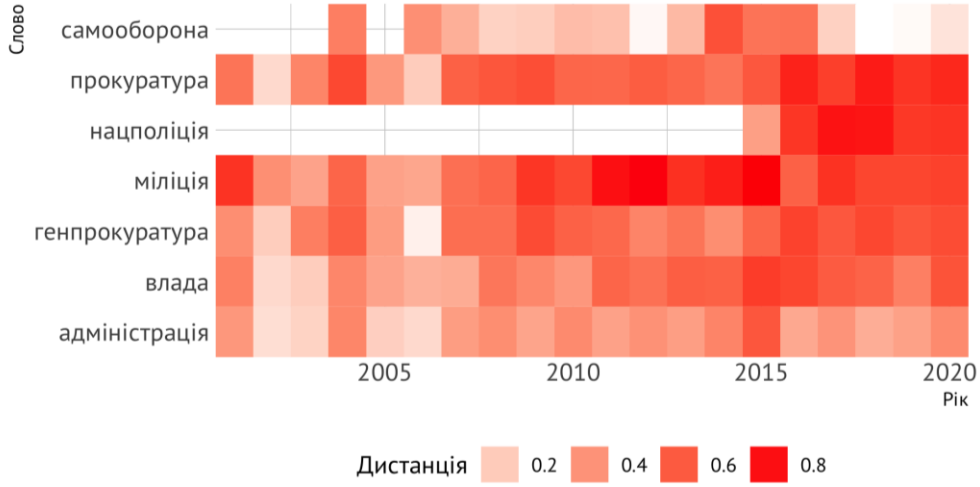

Рис. 2. *Найбільш семантично близькі слова до слова «поліція» за роками*
Джерело: авторські дані.

Отже, ми бачимо, що векторне представлення слів є хорошим методом для дослідження семантики слів. Цей метод продовжує розвиватися, існують способи побудови векторного представлення відразу для цілих текстів, які теж можна використати для соціологічного аналізу.

2.2. Моделювання тем *(topic modeling)*

Моделювання тем на сьогодні є найпопулярнішим методом обчислювального аналізу тексту, що застосовується в соціологічних дослідженнях. Зокрема, це дослідження Паоло Ді Маджіо «Використання спорідненості між моделюванням тем та соціологічною перспективою культури: застосування до висвітлення газет урядового фонду мистецтв США», де науковець намагався за допомогою цього аналізу визначити відмінності між газетами в плані висвітлення культури (DiMaggio et al., 2013).

Ліндштедт у статті «Структурне моделювання тем для соціальних науковців: коротке кейс-стаді з літератури про соціальні рухи за 2005–2017 роки» за допомогою методу латентного розміщення Діріхле визначив топ-24 теми, про які писали соціальні науковці в цей період (Lindstedt, 2019).

Ротшильд, Хават, Шафранек і Басбі у статті «Голубині партійці: стереотипи прихильників партій та партійна поляризація» використовували метод структурного тематичного моделювання для виокремлення стереотипів про демократів та республіканців у США. Таке моделювання вони здійснювали на основі аналізу відповідей респондентів на прохання описати представників партії, яку ці респонденти не підтримували (Rothschild et al., 2019).

Моделювання тем належить до завдань навчання без учителя. Серед методів моделювання тем текстів найпопулярніше латентне розміщення Діріхле (LDA) (Blei et al., 2003). Також цікавим є метод імовірнісного латентно-семантичного аналізу (pLSA) (Hofmann, 1999), однак він повільніший у розрахунку. Ці методи постали до появи векторного представлення слів, тому спираються на мішок слів, який не відображає змістові відносини між окремим словами, що дещо обмежує їх потенціал. Для експерименту ми використали модифіковану версію цього методу під назвою *Top2Vec*[1], яка застосовує векторне представлення слів і документів (Angelov, 2020). Перевагою останнього методу є його непараметричність. LDA та pLSA вимагають від дослідників, які їх використовують, задати кількість тем, котрі модель має згенерувати, тоді як Top2Vec самостійно визначає їх кількість[2].

[1] Його технічна імплементація. URL: https://github.com/ddangelov/Top2Vec

[2] Для цього n-мірне векторне представлення слів і документів зменшується до двох вимірів через UMAP. На наступному етапі це двовимірне представлення піддається кластерному аналізу методом HDBSCAN.

Ми використали цей підхід для пошуку тем у новинах «Української правди», де фігурувала поліція. Нам удалося загалом побудувати модель із 1500 темами розміром від 17 до 4500 документів. Серед них виокремили топ-5 найбільш асоційованих із поліцією тем[1].

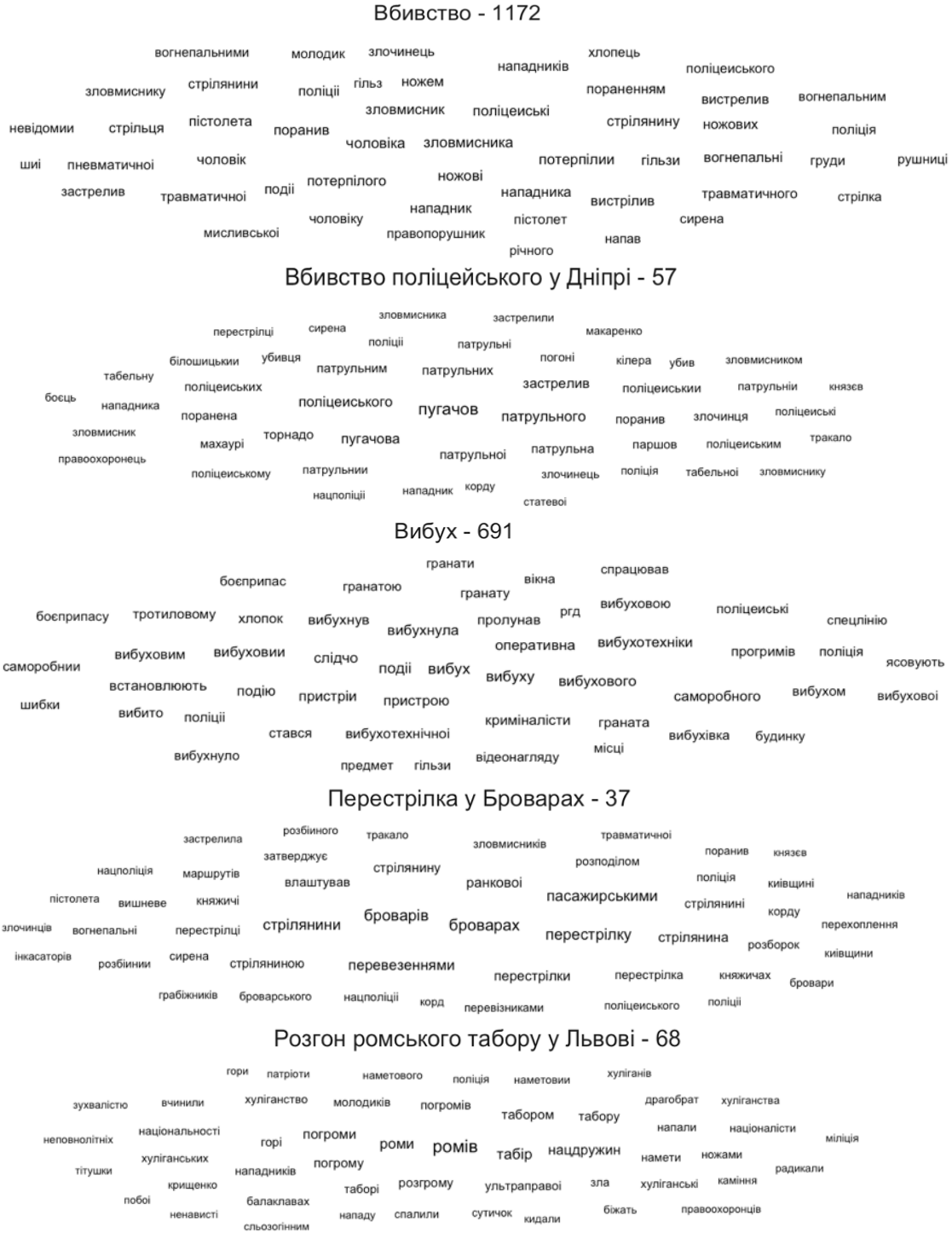

Рис. 3. *Топ-5 тем, де фігурувала поліція*
Джерело: авторські дані.

[1] Особливістю всіх методів моделювання тем є те, що кожна тема – це, по суті, набір із слів, які найбільше з нею асоціюються, тому, коли говоримо, «найбільш асоційована зі словом тема», – це означає, що вона входить у топ-50 асоційованих із темою слів.

Із представлених вище наборів ключових слів із теми можна чітко відгадати, про яку саме тему йдеться. У трьох випадках це резонансні події (убивство поліцейських у Дніпрі, погром табору ромів у Львові та перестрілка в Броварах), у двох інших модель створила збірні теми, де об'єднала новини, пов'язані з убивствами й нападами, та новини про вибухи.

Отже, ми бачимо, що сучасні підходи до моделювання тем дають змогу створити на основі масиву документів легко інтерпретовані тематики.

2.3. Аналіз тональностей *(sentiment analysis)*

Аналіз сентиментів у тексті також є популярним дослідницьким методом.

Флорес у статті «Чи формують антиіміграційні закони суспільні настрої? Дослідження Арізони з використанням даних Twitter» використовує для аналізу тональності твіттер-повідомлень про мігрантів словники тональностей (Flores, 2017).

Існує низка підходів до аналізу тональностей, які можна умовно поділити на дві групи: 1) словниковий підхід (тут дослідники беруть наявні словники тональностей для слів, де позначено рівень сентиментів, ці значення потім сумуються для кожного аналізованого тексту); 2) навчання з учителем (є корпус текстів із попередньо вручну закодованими сентиментами, на основі яких вивчається модель машинного навчання, яка відгадує сентименти на інших текстах).

Досвід застосування другого підходу є на українських і російських даних ЗМІ. В. Бобічев, О. Каніщева та О. Чередниченко в статті «Сентимент аналіз в українських і російських новинах» зробили навчання різних моделей з учителем (насамперед SVM) та проаналізували показники якості навчених моделей на метриці F1 (Bobichev et al., 2017).

Ми застосували обидва підходи для виявлення динаміки оцінки поліції в новинах «Української правди».

Для реалізації першого взято словник тональностей, створений волонтерами проєкту lang.org.ua. Цей словник містить 3442 слів, для яких методом усереднених експертних оцінок проставлено тональності від -2 (сильний негатив) до 2 (сильний позитив) (Шеховцов).

Другий підхід ми реалізували через побудови моделі логістичної регресії для прогнозування наявності негативу в повідомленні на основі закодованих tf-idf кодуванням новинних повідомлень. Вибір логістичної регресії зумовлений тим, що ми маємо справу із завданням бінарної класифікації (потрібно навчити модель прогнозувати два класи – негативний і позитивний сентименти). Логістична регресія ґрунтується на лінійній логіці побудови моделі, тому її важко перенавчити під дані, на яких вона навчається. Відтак для цього методу характерна висока надійність результатів прогнозу.

Таблиця 2

Динаміка тональності повідомлень про поліцію за роками

Рік	Навчання з учителем	Словник
2001	0.0885812	-0.0155637
2002	0.0871971	-0.0137107
2003	0.0815680	-0.0155763
2004	0.0893810	-0.0090909
2005	0.0915999	-0.0134529
2006	0.0949106	-0.0098771
2007	0.0943140	-0.0075758
2008	0.0912443	-0.0133586
2009	0.0956697	-0.0115607
2010	0.0963092	-0.0099010
2011	0.0931102	-0.0106955
2012	0.0955100	-0.0101266
2013	0.0938035	-0.0122959
2014	0.0923548	-0.0088235
2015	0.0944934	-0.0083683
2016	0.0935684	-0.0104167
2017	0.0929389	-0.0116279
2018	0.0927650	-0.0138889
2019	0.0941034	-0.0138889
2020	0.0946125	-0.0132450

Оцінка поліції всі роки стабільна й мало відхиляється від нейтрального значення 0. Це пов'язано з тим, що до аналізу брали саме новинні повідомлення, які за дотримання журналістами професійних стандартів (а саме утримання в повідомленнях від оцінних суджень, які більше властиві для персональних блогів чи розлогих статей) повинні мати нейтральне забарвлення. Попри наші очікування, нейтральний тон повідомлень зберігся й у 2013–2014 рр., коли дії спецпідрозділів поліції проти протестувальників, на нашу думку, могли вилитись у їх негативне висвітлення в ЗМІ.

Отже, аналіз тональності документів може допомогти нам зрозуміти, чи намагались автори документів показати щось у негативному або позитивному світлі. У випадку аналізу висвітлення дій поліції автори новин не робили таких спроб, дотримуючись при цьому журналістських стандартів. Але ми могли виявити їх порушення за допомогою цього методу. Не виключаємо, що обрані нами методи можуть не підходити для аналізу тональностей новинних повідомлень (інші дослідники переважно тестували свої моделі оцінки тональностей на датасетах із Twitter-повідомленнями чи відгуками на сайтах, у яких писати багато оцінних суджень вважається нормою).

ВИСНОВКИ

Нові методи автоматизованої обробки природної мови значно розширюють можливості соціологічного аналізу документів. Це стало можливим через те, що ці методи дають змогу фіксувати семантику текстів, змістову близькість слів, яка не виражається просто в їх схожому написанні. Також усе менше завдань обробки потребують попереднього кодування масивів для навчання моделей. Ще одним важливим проривом є те, що застосування цих методів для аналізу документів робить його масштабованим: для машини немає принципової різниці проаналізувати десять документів чи один мільйон, тоді як при ручному кодуванні складність аналізу значною мірою залежить від розміру корпусу текстів, який планується проаналізувати. Це також усуває проблему вибірки в дослідженнях за допомогою документів, оскільки автоматизований аналіз дає змогу використати суцільну вибірку. Звісно, методи обробки природної мови задумувалися насамперед для розв'язання несоціологічних завдань, тому існує необхідність у їх переосмисленні для того, аби уможливити соціологічний аналіз із їх використанням. На сьогодні ми виокремили три класи завдань, де вже ці методи успішно застосовують, і на прикладі аналізу масиву новин інтернет-видання «Українська правда» показали, як це можна зробити. Це завдання зміщення семантичного значення слів, моделювання тем та сентимент аналізу. У першому випадку виявили, що поліція в останні роки сильніше ототожнюється з концептом влади, ніж раніше. У другому випадку виявлено п'ять найбільш резонансних тем, пов'язаних із діяльністю поліції. І в третьому випадку встановлено, що новинні повідомлення про поліцію стабільно мають нейтральну тональність.

Поряд із перевагами в автоматизованого аналізу документів є й недоліки. Головні з них – те, що виконання кожної технічної задачі (наприклад класифікації) потребує навчання нової моделі, заточеної під це завдання. Для навчання таких моделей потрібні масиви текстів, розмічених дослідниками. Це також створює другий недолік – моделі мають не 100 % точність роботи й дуже залежать від якості даних й алгоритму, за допомогою яких були навчені. Третій недолік – технічні ресурси, що потрібні для обробки текстів. Частина популярних нині методів ґрунтується на алгоритмах нейронних мереж, які містять мільйони параметрів. Для навчання таких моделей прийнято використовувати менш доступні графічні процесори.

Отже, нам удалося систематизувати наявні нині знання про використання методів обробки природної мови в соціології.

ДЖЕРЕЛА ТА ЛІТЕРАТУРА

Shekhovtsov, S., Chaplynskyi, D., Petriv, O. Tonal dictionary of the Ukrainian language. Retrieved March 28, 2021 from https://lang.org.ua/uk/dictionaries/

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. *arXiv*. Retrieved August 19, 2020 from http://arxiv.org/abs/2008.09470

Bamler, R., Mandt, S. (2017). Dynamic word embeddings. *34th International Conference on Machine Learning, ICML*, 2017, 1, 607–621. Retrieved August 19, 2020 from http://arxiv.org/abs/1702.08359

Blei, D. M., Ng, A. Y., Edu, J. B. (2003). *Latent Dirichlma inlocation Michael I. Jordan*, Jan; 3, 993–1022.

Bobichev, V., Kanishcheva, O., Cherednichenko, O. (2017). Sentiment analysis in the Ukrainian and Russian news. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON). doi: 10.1109/ukrcon.2017.8100410

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. Retrieved May 24, 2019 from http://arxiv.org/abs/1810.04805

Di Carlo, V., Bianchi, F., Palmonari, M. (2019). Training Temporal Word Embeddings with a Compass. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 6326–6334. doi: 10.1609/aaai.v33i01.33016326

DiMaggio, P., Nag, M., Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Flores, R. D. (2017). Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 Using Twitter Data. *American Journal of Sociology*, 123(2), 333–384. https://doi.org/10.1086/692983

Harris, Z. S. (1954). Distributional structure, 10(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hofmann, T. (1999). Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR '99. doi: 10.1145/312624.312649

Kozlowski, A. C., Taddy, M., Evans, J. A. (2018). The Geometry of Culture: Analyzing Meaning through Word Embeddings. *American Sociological Review*, 84(5), 905–949. https://doi.org/10.1177/0003122419877135

Lemke, M., Wiedemann, G. (2016). *Text mining in den sozialwissenschaften*. Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-07224-7

Lindstedt, N. C. (2019). Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017. *Social Currents*, 6(4), 307–318. https://doi.org/10.1177/2329496519846505

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). *Efficient estimation of word representations in vector space*. Retrieved May 22, 2019 from http://arxiv.org/abs/1301.3781

Pennington, J., Socher, R., Manning, C. (2014). Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). doi:10.3115/v1/d14-1162

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. *Language Models are Unsupervised Multitask Learners*. Retrieved January 1, 2020 from https://github.com/codelucas/newspaper

Rothschild, J. E., Howat, A. J., Shafranek, R. M., Busby, E. C. (2019). Pigeonholing Partisans: Stereotypes of Party Supporters and Partisan Polarization. *Political Behavior*, 41(2), 423–443. https://doi.org/10.1007/s11109-018-9457-5

Stone, P. J., Dunphy, D. C., Smith, M. S., Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017–December*, 5999–6009. Retrieved December 6, 2017 from http://arxiv.org/abs/1706.03762

Yao, Z., Sun, Y., Ding, W., Rao, N., Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. *WSDM 2018 – Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, 2018–Febua, 673–681. https://doi.org/10.1145/3159652.3159703

Yin, W., Kann, K., Yu, M., Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *CoRR*, abs/1702.01923. Retrieved February 7, 2017 from http://arxiv.org/abs/1702.01923

Zhang, H. (2019). Dynamic Word Embedding for News Analysis. *UCLA*. ProQuest ID: Zhang_ucla_0031N_18000. Merritt ID: ark:/13030/m5wh7p2f. Retrieved January 1, 2020 from https://escholarship.org/uc/item/9tp9g31f