УДК 311

АНАЛІЗ НОВИН НА ГЕНДЕРНУ ТЕМАТИКУ ЗАСОБАМИ АЛГОРИТМІВ ТЕКСТОВОГО АНАЛІЗУ

Віктор Туртуріка

Наукові керівники: кандидат фізико-математичних наук, доцент кафедри математики та цифрових технологій Акбаш К.С., кандидат фізико-математичних наук, доцент кафедри математики та цифрових технологій Макарчук О. П.

Центральноукраїнський державний університет імені Володимира Винниченка

Анотація. *Стаття присвячена дослідженню можливостей застосування алгоритмів текстового аналізу для вивчення новин на гендерну тематику. Розглядаються методи частотного аналізу слів, виявлення ключових термінів за допомогою алгоритму TF-IDF та аналізу емоційного забарвлення текстів для ідентифікації основних тенденцій і стереотипів, пов'язаних із гендерними питаннями.*

Ключові слова: *текстовий аналіз, гендерна рівність, частотний аналіз, емоційний аналіз, алгоритми TF-IDF.*

ANALYSIS OF NEWS ON GENDER THEME USING TEXT ANALYSIS ALGORITHMS

Viktor Turturika

Academic supervisors: Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Mathematics and Digital Technologies K.S.Akbash; Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Mathematics and Digital Technologies O. P.Makarchuk

Volodymyr Vynnychenko Central Ukrainian State University

Annotation. *The article is devoted to researching the possibilities of using text analysis algorithms to study news on gender topics. The methods of frequency analysis of words, identification of key terms using the TF-IDF algorithm and analysis of emotional coloring of texts to identify the main trends and stereotypes related to gender issues are considered.*

Keywords: *text analysis, gender equality, frequency analysis, emotional analysis, TF-IDF algorithms.*

Актуальність роботи.

У сучасному інформаційному просторі, де обсяг текстових даних постійно зростає, текстовий аналіз є важливим інструментом для узагальнення та інтерпретації великих обсягів інформації. Це особливо актуально в гендерних дослідженнях, які вимагають обробки даних із різних джерел, таких як медіа, соціальні мережі, законодавчі акти та наукові публікації. Аналіз текстів на гендерну тематику дозволяє дослідникам виявляти суспільні установки, стереотипи та тренди, а також оцінювати рівень уваги до питань гендерної рівності.

Результати такого аналізу можуть бути застосовані для вдосконалення освітніх програм, створення рекомендацій у сфері державної політики та покращення медійних стандартів. Це підкреслює важливість текстового аналізу як ключового інструменту в гендерних дослідженнях, спрямованих на досягнення рівноправності та гармонії в суспільстві.

Серед технік, що використовуються у гендерному аналізі, важливе місце займає частотний аналіз і метод TF-IDF, який допомагає виявити найбільш релевантні слова й фрази, що асоціюються з гендерними питаннями. Ці техніки дозволяють ідентифікувати ключові теми, а також проаналізувати, як часто й у якому контексті обговорюються питання гендерної рівності. Методи інтелектуального аналізу також дозволяють проводити тональність і настрій текстів, що допомагає зрозуміти емоційне забарвлення контенту, пов’язаного з гендерними питаннями. Таким чином, інтелектуальний аналіз текстових даних відкриває нові можливості для гендерного дослідження, дозволяючи не лише якісно оцінити тематику, а й кількісно виміряти її аспекти, що сприяє більш глибокому розумінню соціальних процесів, пов'язаних із гендерною рівністю [2]

Метою статті є застосування алгоритмів для аналізу даних текстового характеру до новинних статей пов’язаних із гендерною тематикою.

Вибір та накопичення даних.

Для аналізу текстів на гендерну тематику було обрано два провідних інформаційних ресурси: сайт державного суспільного мовника *Суспільне Новини* [6] (новини за тегом «Гендерна рівність») та сайт *Радіо Свобода* [5] (новини, в яких згадується тематика гендеру). Ці ресурси висвітлюють важливі соціальні та культурні питання, зокрема теми рівності, гендерних стереотипів, прав жінок. Завдяки цьому вибору можна отримати репрезентативну базу даних для аналізу різноманітних аспектів гендерної тематики в українських і міжнародних медіа.

Із сайту *Суспільного* було зібрано 168 статей, які публікувалися в період із 2020 по 2024 роки. Тексти охоплюють питання гендерної рівності, прав жінок, дискримінації, а також відображають новітні суспільні ініціативи і зміни у законодавстві, пов'язані з гендером. Зосередження на останніх роках дає змогу дослідити актуальні тенденції та новітній підхід до гендерних питань, характерний для сучасних українських медіа [6].

Сайт *Радіо Свобода* надає матеріали з 2008 по 2024 рік, загалом 364 статі, які містять згадки про гендерну тематику. Такий довготривалий часовий період дозволяє відстежити динаміку змін у висвітленні гендерних питань протягом більш ніж десятиліття, а також порівняти ранні підходи до теми з сучасними. Теми включають гендерну нерівність, боротьбу за права жінок, а також реакцію суспільства на різні аспекти гендерних питань. Окрім цього, матеріали з *Радіо Свобода* надають можливість розглянути міжнародний контекст у висвітленні гендерної тематики [5].

Для обох сайтів (*Суспільного* та *Радіо Свобода*) було виконано процес агрегування, що складався з двох основних етапів:

– *Формування списку статей* – На першому етапі було здійснено обхід загального списку статей на кожному сайті, у межах якого зібрано URL-посилання на кожну статтю, що відповідає вибраним темам (гендерна рівність, гендерна дискримінація тощо).

– *Завантаження текстового вмісту статей*. На другому скрипти реалізовані мовою R завантажують тексти статей, їх заголовки, посилання та

дати публікації, зчитуючи посилання з відповідних файлів. Завантажені дані зберігаються у *csv* файлах, кожен з яких має структуру:

- *date_published* — дата публікації статті,
- *link* — URL-адреса на статтю,
- *title* — заголовок статті,
- *text* — текст статті.

Попередня підготовка даних.

Підготовка текстових даних є ключовим етапом перед проведенням текстового аналізу, оскільки забезпечує точність і релевантність результатів. Застосовані методи включають *стемінг*, *вилучення стоп-слів*, *видалення спеціальних символів* та приведення тексту до *нижнього регістру.*

Стемінг – зведення слів до їх базових форм дозволяє об'єднати різні варіації одного терміна, знижуючи кількість варіантів одного й того ж слова. Це суттєво покращує точність лексичного аналізу, зокрема для української мови, де слова можуть змінюватися в залежності від відмінків та часу. Оскільки готові бібліотеки R, такі як *tm* або *SnowballC*, орієнтовані на англійську мову, для українських текстів було розроблено власну версію алгоритму Портера для стемінгу, адаптовану до специфічних суфіксів та морфології українських стоп-слів. Щоб зменшити інформаційний шум, з текстів було вилучено *стоп-слова* [32] — це поширені службові слова, які не мають значущого змісту у контексті тематичного аналізу. Вилучення стоп-слів дозволяє зосередитись на ключових словах і поняттях, що покращує якість подальшого аналізу. Для цього було використано український список стоп-слів, який враховує частотні та службові слова.

Видалення спеціальних символів та *зміна регістру.* Спеціальні символи, такі як розділові знаки, лапки чи додаткові пробіли, були видалені для створення уніфікованого формату тексту. Приведення тексту до нижнього регістру забезпечує однорідність написання слів і знижує ризик дублювання понять через різні варіанти регістру (наприклад, "гендер" та "Гендер").

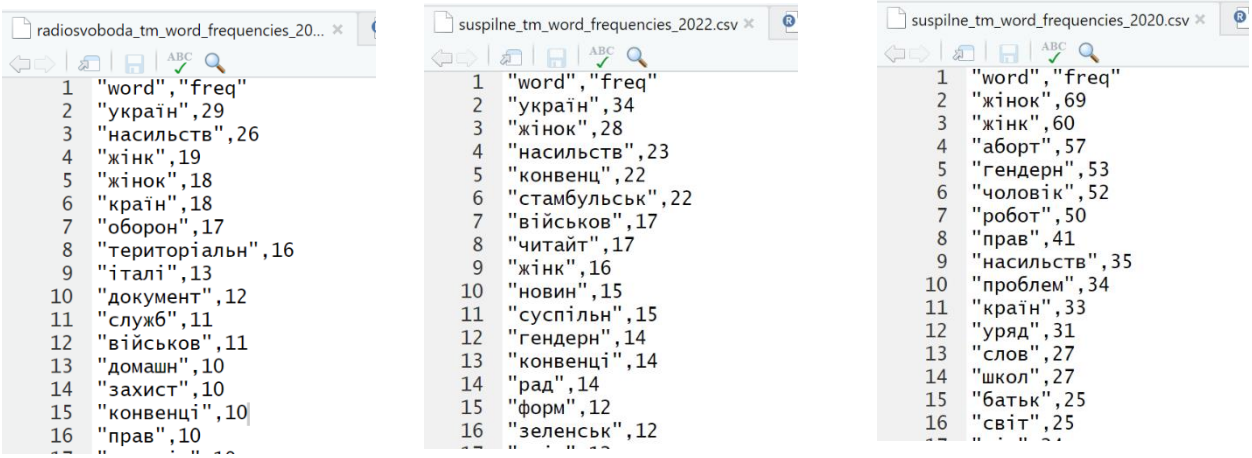
В подальших викладках пам’ятатимемо, що всі вище перелічені дії було виконано і немає сенсу знову коментувати їх.

Частотний аналіз із використанням пакету tm

Для реалізації одного з найпростіших алгоритмів – частотного аналізу на наших даних для дослідження будемо використовувати засоби пакету *tm*. Проаналізуємо, які слова зустрічаються найчастіше у текстах новин пов’язаних із гендерною тематикою у статтях *Суспільного* та *Радіо Свобода*. Розрахуємо загальну частоту та частоту за роками (для *Суспільного* із 2020 по 2024 роки, та із 2008 по 2024 рік для *Радіо Свобода* відповідно).

Результати представимо у вигляді CSV файлів де кожному Також, сформуємо додаткові файли із відфільтрованими за роками даними. Окрім того, корисно візуалізувати дані, використовуючи пакет *ggplot* [3]. Будемо дивитись яка десятка слів є найпоширенішою в загальному корпусі та по роках

Будемо порівнювати результати по кожному із новинних сайтів та враховувати рік випуску тієї чи іншої статті. Так, для 2022 року, очікувано, найбільш вживаними термінами були слова пов’язані із повномасштабним вторгненням – *Україна, насильство, конвенція, оборона, військові* (Рис. 1, а-б).



а) б) в)

Рис. 1. Найбільш вживані терміни за 2022 (а-б) та за 2020 (в) роки

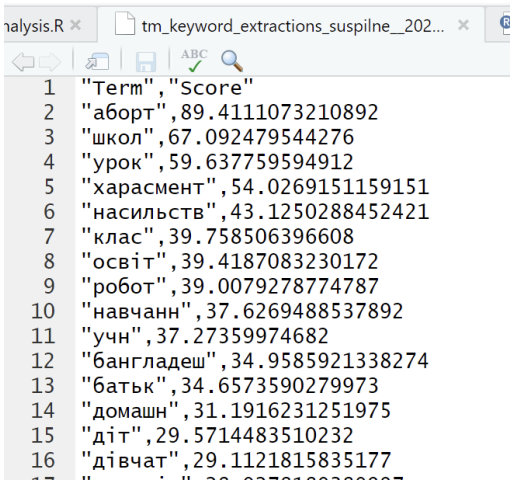
Оцінимо, які ключові слова переважали у роки до повномасштабного вторгнення. Для цього розглянемо результати за 2020 рік від Суспільного (Рис 1.в). Так, для на Суспільному найчастіше згадуються теми *абортів*, *гендеру*, *жінок* та *чоловіків* (що очевидно), *прав*, *свобод* та *проблем насильства*. Тим часом для Радіо Свобода найбільш вживаними термінами у 2018 році були *права*, *свободи* та *рівність.*

Виокремлення ключових слів із використанням пакету tm.

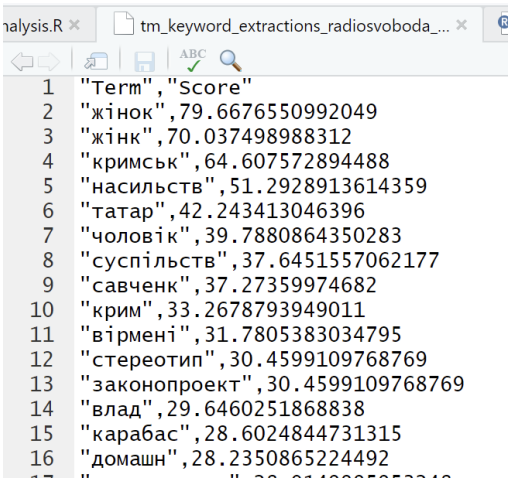
Будемо шукати ключові слова за допомогою алгоритму *TF-IDF*, який дозволяє визначити найбільш значущі терміни у текстовому документі. Алгоритм обчислює важливість кожного слова на основі його частоти в конкретному документі та частоти в усіх документах корпусу [2]. Це допомагає виявити терміни, які є специфічними для даного документа і менш поширеними в інших, тим самим вказуючи на ключові слова, що найкраще відображають його зміст.

Подібно до попереднього прикладу оформимо результати у csv файли та сформуємо діаграми для 50 слів які мають найбільший бал *TF-IDF* по загальному корпусу текстових даних, згрупованих по роках та новинних сайтах. Для графіків сформуємо діаграми по 50 слів які мають найбільший бал *TF-IDF* як по загальному корпусу текстових даних, так і згрупованих по роках. Ці терміни будуть представляти окремі теми новинних статей

Наприклад розглянемо ключові теми на Суспільному у 2020 році (Рис 2.а). Бачимо, що в цьому році найбільше піднімались теми *абортів* (частотний аналіз також підтверджує цей факт), *шкіл*, *уроків*, *харасменту*, *освіти, роботи* та *дітей*. Водночас у 2016 році на Радіо Свобода поряд із темами *жінок*, *насильства*, *чоловіків* та *стереотипів* піднімались теми *кримських татар* та згадувалась *Надія Савченко* (Рис 2.б).



1	"Term","Score"
2	"аборт",89.4111073210892
3	"школ",67.092479544276
4	"урок",59.637759594912
5	"харасмент",54.0269151159151
6	"насильств",43.1250288452421
7	"клас",39.758506396608
8	"освіт",39.4187083230172
9	"робот",39.0079278774787
10	"навчанн",37.6269488537892
11	"учн",37.27359974682
12	"бангладеш",34.9585921338274
13	"батьк",34.6573590279973
14	"домашн",31.1916231251975
15	"діт",29.5714483510232
16	"дівчат",29.1121815835177



1	"Term","Score"
2	"жінок",79.6676550992049
3	"жінк",70.037498988312
4	"кримськ",64.607572894488
5	"насильств",51.2928913614359
6	"татар",42.243413046396
7	"чоловік",39.7880864350283
8	"суспільств",37.6451557062177
9	"савченк",37.27359974682
10	"крим",33.2678793949011
11	"вірмені",31.7805383034795
12	"стереотип",30.4599109768769
13	"законопроект",30.4599109768769
14	"влад",29.6460251868838
15	"карабас",28.6024844731315
16	"домашн",28.2350865224492

а) б)

Рис. 2. Ключові теми статей на Суспільному у 2020 році (а) та на Радіо Свобода у 2016 році (б)

Далі розглянемо тематику новин на Радіо Свобода у 2023 році. Поряд із очевидними темами *війни*, *насильства*, *ворогів* та *війська*, спостерігаємо згадку про *жінок* (очевидно), *законопроекти*, *фемінізм* та *квоти.*

Аналіз емоційної забарвленості за допомогою пакету tidytext

Для проведення аналізу статей на емоційну забарвленість скористаємось можливостями пакету *tidytext,* а саме особливостями формату *tidy* який підтримує операції *inner_join*, *distinc* та *left_join*. Єдиним (і не зовсім суттєвим) недоліком використання *tidytext* для цієї задачі є відсутність лексиконів для української мови. Однак, у відкритих джерелах можна знайти відповідні лексикони, попередньо перетворивши коефіцієнти емоційного забарвлення до формату, який розумітиме *tidytext.*

Представимо результати у вигляді csv файлів, який міститиме заголовок статті, негативну оцінку, позитивну оцінку та остаточну оцінку статті.

Проаналізуємо деякі статті та отриману оцінку емоційного забарвлення. Так, виберемо кілька результатів для Суспільного (Табл. 1):

Таблиця 1. Оцінка емоційного забарвлення деяких новин Суспільного у 2024 році

Заголовок	Негативна лексика	Позитивна лексика	Оцінка
"Міноборони впроваджує стандарти НАТО щодо гендерної рівності у ЗСУ"	0	3	3
"Рада прийняла за основу законопроєкт щодо посилення соцзахисту військових. Окремим пунктом виділено недопущення сексизму"	0	2	2
"В Афганістані таліби виключили зі шкіл дівчат",4,2,-2	4	2	-2
"В Іспанії чоловіки-вчителі на знак протесту прийшли на роботу у спідницях"	5	3	-2

Дійсно, новини які описують покращення та якісні зміни (новини 1,2 із Таблиці 1) алгоритм визначає як статті із позитивним емоційним контекстом. Водночас новини (3-4 із таблиці 1) які висвітлюють заворушення або грубе порушення прав людини відмічені як негативні. Звичайно, що оцінка не абсолютно вірною для всіх новин із досліджуваного набору, проте більшість результатів вірно інтерпретують емоційне забарвлення текстів. Для покращення результатів рекомендовано збільшити та покращити набір лексем.

Висновки та подальші перспективи.

Аналіз текстових даних є важливим та необхідним інструментом у сучасному світі, де обсяг текстової інформації постійно зростає. Застосування класичних статистичних методів у поєднанні з сучасними алгоритмами машинного навчання забезпечує не тільки високу точність аналізу, але й можливість інтерпретації результатів. Це робить їх надзвичайно корисними для вирішення широкого спектру завдань у різних галузях.

Аналіз текстових даних новин на гендерну тематику дозволяє виявити ключові теми та тенденції у висвітленні питань, пов'язаних із гендерною рівністю. Використання методів аналізу тексту допомагає визначити, які аспекти гендерної тематики найбільш обговорювані, а також відслідковувати

зміни в акцентах і тональності з часом. Завдяки алгоритмам обробки природної мови можна ідентифікувати характерні слова та фрази, що відображають загальні настрої та бачення гендерних питань у медіа, що сприяє більшій об'єктивності та повноті аналізу текстів. Результати дослідження можуть бути застосовані для аналізу соціальних настроїв, оцінки популярності тем та прийняття обґрунтованих рішень у різних галузях.

Подяки. Стаття підготовлена в рамках проєкту ERASMUS-JMO-2021-HEI-TCH-RSCH "Субнаціональна гендерна рівність: баланс цінностей ЄС та українських реалій" № 101047451. Проєкт фінансується за підтримки Європейської комісії. Ця публікація відображає лише погляди авторів. Єврокомісія не несе відповідальності за будь-яке використання інформації, що міститься в ній.

Список використаної літератури

1. Akbash K., Pasichnyk N., Rizhniak R. Generalization of calculation methods for gender indices in demographic and social statistics. Regional Statistics, Vol. 8. No. 2. 2018: 170183; DOI: 10.15196/RS080205 http://www.ksh.hu/docs/hun/xftp/terstat/2018/rs080205.pdf
2. Silge J. Text Mining with R / J. Silge, D. Robinson., 2017. – 190 с.
3. Text Mining: Term vs. Document Frequency [Електронний ресурс] – Режим доступу до ресурсу: https://uc-r.github.io/tf-idf_analysis
4. Глибовець А. Алгоритм токенізації та стемінгу для текстів українською мовою / А. Глибовець, В. Точицький. – 2017.
5. Радіо Свобода [Електронний ресурс] – Режим доступу до ресурсу: https://www.radiosvoboda.org/s?k=гендерна+рівність&tab=news&r=any&pp=50
6. Суспільне Новини [Електронний ресурс] – Режим доступу до ресурсу: https://suspilne.media/tag/genderna-rivnist/

Відомості про автора:

Туртуріка Віктор Ігорович – студент ІІ курсу магістратури факультету математики, природничих наук та технологій Центральноукраїнського державного університету імені Володимира Винниченка, тел. +380 66 504 12 61, e-mail: v.turturika@gmail.com