УДК 004.5 + 004.89

Д.П. Урлапова

О.В. Бісікало

АНАЛІЗ МОВНИХ МАРКЕРІВ У СТЕНОГРАМАХ ВЕРХОВНОЇ РАДИ УКРАЇНИ У ВИГЛЯДІ ВЕБДОДАТКА НА PYTHON ТА STREAMLIT

Вінницький національний технічний університет

Анотація

Дослідження присвячено розробленню вебдодатка для аналізу мовних маркерів у стенограмах Верховної Ради України. У роботі використано офіційні тексти пленарних засідань за 2023–2026 роки, виконано їх очищення від технічного та процедурного шуму, сформовано тематичні групи та реалізовано підрахунок частотності мовних маркерів. Програмний засіб створено мовою Python із використанням бібліотек для обробки тексту, візуалізації даних і фреймворку Streamlit. Отримані результати дали змогу виявити зміну риторичних акцентів у парламентському дискурсі за роками та подати їх у вигляді таблиць і графіків. Практична цінність роботи полягає у створенні зручного інструмента для подальшого аналізу великих українськомовних текстів.

Ключові слова: *мовні маркери; стенограми Верховної Ради України; політичний дискурс; Python; Streamlit; тематичний аналіз; вебдодаток.*

Abstract

The study is devoted to the development of a web application for analyzing language markers in the transcripts of the Verkhovna Rada of Ukraine. The research uses official plenary session transcripts for 2023–2026, performs their cleaning from technical and procedural noise, forms thematic groups, and implements frequency analysis of language markers. The software tool was developed in Python using text processing and data visualization libraries together with the Streamlit framework. The obtained results made it possible to identify changes in rhetorical emphases in parliamentary discourse over the years and present them in the form of tables and charts. The practical value of the work lies in creating a convenient tool for further analysis of large Ukrainian-language texts.

Keywords: *language markers; transcripts of the Verkhovna Rada of Ukraine; parliamentary discourse; thematic groups; web application; Python; Streamlit; text analysis; data visualization.*

Вступ

Розвиток відкритих цифрових ресурсів органів державної влади створює нові можливості для автоматизованого аналізу політичного мовлення. Стенограми парламентських засідань становлять цінний текстовий матеріал, оскільки дають змогу простежити, які проблеми виходять на перший план у публічному обговоренні, як змінюються риторичні акценти та які тематичні напрями домінують у різні періоди.

Для сучасних досліджень це особливо важливо, адже методи комп'ютерного аналізу тексту вже застосовуються до парламентських дебатів і дають змогу виявляти тематичні та позиційні зміни в політичному дискурсі [1, 2]. Для українського контексту така задача набуває додаткової актуальності в умовах повномасштабної війни, коли в публічному порядку денному одночасно присутні безпековий, економічний, соціальний та відновлювальний виміри [3, 4].

Отже, у цьому напрямку вже доведено ефективність автоматизованого аналізу парламентських текстів, однак прикладні засоби для роботи саме з українськомовними стенограмами Верховної Ради України у доступному вебформаті представлені обмежено.

Метою дослідження є розроблення вебдодатка для автоматизованого аналізу мовних маркерів у стенограмах Верховної Ради України, який забезпечує очищення текстів, виділення тематичних груп, агрегування результатів за роками та їхню візуалізацію.

Результати дослідження

Основу для дослідження становлять відкриті стенограми пленарних засідань Верховної Ради України з офіційного розділу «Стенограми» на вебпорталі парламенту [5]. До початкового корпусу включено по два тексти для кожного року 2023, 2024, 2025 і 2026, що дало змогу отримати річні зрізи для первинного порівняльного аналізу.

Програмний засіб реалізовано мовою Python із використанням бібліотеки pandas для табличного опрацювання результатів і фреймворку Streamlit для побудови вебінтерфейсу [6, 7]. Для візуалізації використано matplotlib. Алгоритм роботи програми охоплює зчитування текстів, очищення від технічного та процедурного шуму, токенізацію, зіставлення слів із тематичними словниками, агрегування показників за роками та формування таблиць і графіків.

На етапі попередньої обробки тексту реалізовано очищення стенограм від технічного та процедурного шуму: часових міток, прізвищ мовців у форматі «ПРІЗВИЩЕ І.О.», службових вставок у дужках, звертань, процедурних реплік і частини стоп-слів. Після очищення виконується токенізація та фільтрація лексем, що дозволяє перейти до змістового аналізу тексту.

Для аналітичної частини сформовано чотири тематичні групи: «Війна і безпека», «Єдність і підтримка», «Відновлення і майбутнє», «Економіка і соціальна сфера». Вибір саме таких груп здійснено на основі поєднання експертного підходу та незалежних аналітичних і соціологічних матеріалів, які фіксують домінування безпекового, соціально-економічного та відновлювального вимірів у публічному порядку денному України [3, 4]. Окреме виділення групи єдності та підтримки зумовлене тим, що в умовах війни консолідаційна риторика виступає самостійним змістовим напрямом парламентського мовлення.

Алгоритм роботи системи складається з таких етапів: зчитування текстових файлів; очищення тексту та видалення шуму; токенізація; віднесення слів до тематичних груп; агрегування частот за роками; побудова таблиць і графіків у вебдодатку. Структурну схему програмного засобу наведено на рисунку 1.

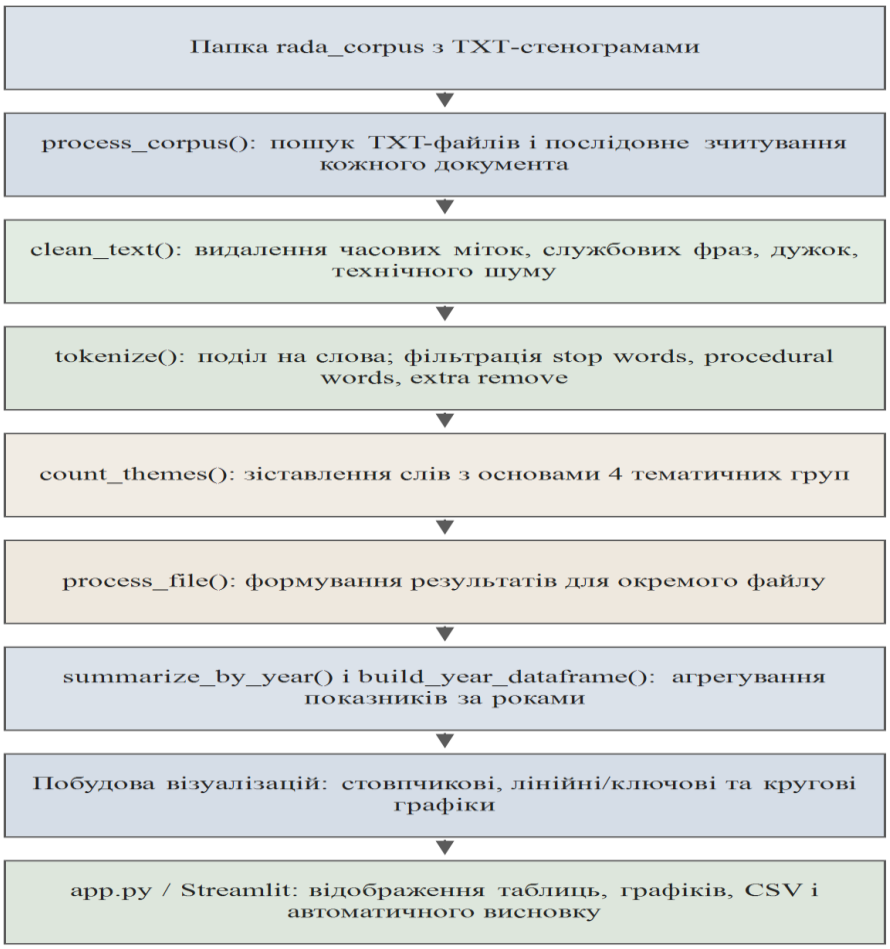

Рис. 1 – Структурна схема роботи програмного засобу

Результати роботи програми було представлено у вебдодатку у вигляді таблиці для окремих файлів, підсумкової таблиці за роками та графіків частоти тематичних груп у розрахунку на 1000 слів. Це забезпечило коректніше порівняння, оскільки тексти стенограм відрізняються за обсягом. Приклад графічного представлення результатів аналізу у вебдодатку наведено на рисунку 2.

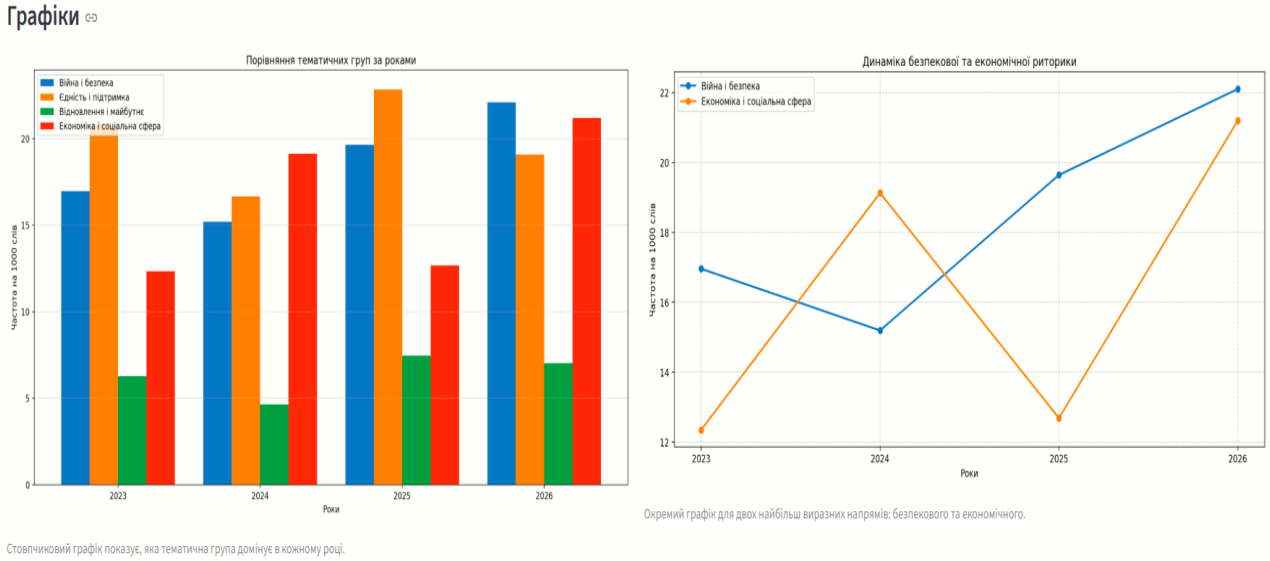

Рисунок 2 – Графічне представлення результатів аналізу мовних маркерів у вебдодатку

Вебдодаток сформував таблицю результатів по окремих файлах, підсумкову таблицю за роками та графіки частоти тематичних груп у розрахунку на 1000 слів. Такий спосіб подання забезпечив коректніше порівняння, оскільки тексти стенограм мають різний обсяг.

Інтерпретація отриманих графіків показала, що у 2023 і 2025 роках виразніше представлена лексика єдності та підтримки; така тенденція узгоджується з аналітичними та соціологічними оцінками, які фіксують високу роль суспільної консолідації, безпеки та солідарності в українському публічному просторі воєнного часу [3, 4]. Для 2024 року помітним є посилення економіко-соціальної тематики, що також відповідає незалежним оцінкам, де серед ключових напрямів суспільного порядку денного виокремлюються питання економічної стійкості, соціальної сфери та відновлення [3, 4]. У 2026 році найбільшою стає частота групи «Війна і безпека», а також зберігається високий рівень економіко-соціальної риторики; це дає підстави пов'язувати домінування відповідних маркерів із переважанням безпекового та соціально-економічного вимірів у публічному порядку денному воєнного періоду [3, 4]. Отже, розроблений засіб дав змогу не лише підрахувати окремі мовні маркери, а й простежити зміну домінантних тематичних акцентів у часовому розрізі.

Практична цінність розробки полягає в тому, що вебдодаток може використовуватися як навчальний інструмент для подальшого аналізу корпусів українськомовних політичних текстів. Корпус стенограм можна розширювати, тематичні словники – уточнювати, а сам підхід – адаптувати до інших типів суспільно значущих текстів.

Висновки

Було запропоновано підхід до аналізу мовних маркерів у стенограмах Верховної Ради України у вигляді вебдодатка на Python та Streamlit. Реалізований програмний засіб забезпечує очищення текстів, виділення тематичних груп, агрегацію результатів за роками та їх візуалізацію у табличній і графічній формах.

Отримані результати підтверджують, що навіть на обмеженому корпусі стенограм автоматизований аналіз дає змогу виявляти зміну риторичних акцентів парламентського дискурсу. Побудовані графіки доцільно інтерпретувати як показники відносної частотності тематичних маркерів у дослідженому корпусі, тобто як індикатори сили риторичного акценту, а не як прогноз політичних подій. Перспективою подальших досліджень є розширення корпусу стенограм,

удосконалення тематичних словників і застосування складніших NLP-підходів для глибшої семантичної інтерпретації.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Abercrombie G., Batista-Navarro R. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review [Electronic resource] // Journal of Computational Social Science. – 2020. – Vol. 3. – P. 245-270. – Available at: https://link.springer.com/article/10.1007/s42001-019-00060-w
2. Ivanusch C. Issue Competition in Parliamentary Speeches? A Computer-based Content Analysis of Legislative Debates in the Austrian Nationalrat [Electronic resource] // Legislative Studies Quarterly. – 2024. – Vol. 49, No. 1. – P. 203-221. – Available at: https://doi.org/10.1111/lsq.12421
3. Разумков Центр. Щомісячний аналітичний огляд «України від війни до миру та відновлення. Аналітичні оцінки» (березень 2024 р.) [Електронний ресурс]. – Режим доступу: https://razumkov.org.ua/statti/shchomisiachnyi-analitychnii-ogliad-ukrainy-vid-viiny-do-myru-ta-vidnovlennia-analitychni-otsinky-berezen-2024r
4. Фонд «Демократичні ініціативи» ім. Ілька Кучеріва. Свобода, безпека, достаток: громадська думка українців під час війни [Електронний ресурс]. – Режим доступу: https://database.dif.org.ua/article/svoboda-bezpeka-dostatok-gromadska-dumka-ukraintsiv-pid-chas-viyni
5. Офіційний вебпортал Верховної Ради України. Стенограми пленарних засідань [Електронний ресурс]. – Режим доступу: https://www.rada.gov.ua/meeting/stenogr
6. pandas documentation. User Guide [Electronic resource]. – Available at: https://pandas.pydata.org/docs/user_guide/
7. Streamlit Documentation. API Reference [Electronic resource]. – Available at: https://docs.streamlit.io/develop/api-reference

Бісікало Олег Володимирович – д-р техн. наук, професор, завідувач кафедри АІІТ, Вінницький національний технічний університет, м. Вінниця, e-mail: obisikalo@vntu.edu.ua

Урлапова Дар'я Павлівна – студентка групи 1іст-23б, факультет автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: dashaurlapova@gmail.com

Bisikalo Oleg V. – Dr.Sc. (Eng.), Professor of the Faculty for Computer Systems and Automatic, Vinnytsia National Technical University, Vinnytsia, e-mail: obisikalo@vntu.edu.ua

Urlapova Daria P. – Department of Automation and intelligent information technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: dashaurlapova@gmail.com