Natural language processing for sociological research with a focus on preprocessing effects in sentiment analysis

Oleksandr Lytvynenko, Myroslava Kukhta and Maksym Yenin[1]

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37 Beresteyskyi Ave., Kyiv, 03056, Ukraine

Abstract

The rapid expansion of digital communication has generated large volumes of unstructured textual data, creating new opportunities and methodological challenges for sociological research. While natural language processing (NLP) tools are increasingly used to analyse public discourse, their effectiveness depends critically on preprocessing strategies that transform raw text into analytically meaningful units. Despite substantial progress in computational linguistics, there is limited empirical evidence on how different preprocessing techniques influence sentiment analysis outcomes in Ukrainian-language sociological datasets. This gap is particularly relevant given the morphological complexity of Ukrainian and the absence of standardised tools optimised for sociological research. This study examines the impact of two preprocessing strategies – lemmatisation and stemming – on sentiment analysis results derived from online news headlines. The dataset comprises 990 headlines scraped from media websites and analysed in a Python environment using spaCy, NLTK, and custom lexicons. Two parallel corpora were constructed: a lemmatised version and a stem-based version, each aligned with a corresponding sentiment dictionary. Sentiment scores were calculated using a modified VADER implementation. A comparative evaluation reveals that 107 out of 990 headlines (10.8%) yield different compound sentiment values solely due to the preprocessing technique. The findings demonstrate that preprocessing is not a purely technical step but a substantive methodological decision that shapes analytical outcomes in computational sociology. The study contributes to the field by providing replicable evidence of how linguistic processing affects sentiment classification in Ukrainian and by outlining a transparent workflow that integrates qualitative interpretive logic with computational efficiency. The results are relevant for sociologists working with digital media, political communication, public opinion monitoring and automated content analysis. Limitations include the experimental nature of the custom sentiment dictionary, noise introduced by OCR, and the constraints of lexicon-based sentiment analysis. These factors suggest directions for future research, including the development of larger Ukrainian sentiment lexicons, the application of transformer-based classifiers and the extension of analysis beyond headlines to full-text corpora.

Keywords

NLP, qualitative data, qualitative research, methodology, sociology, scraping, text classification, sentiment analysis

1. Introduction

Modern sociology increasingly operates in research settings shaped by the demand for measurable, statistically validated evidence. Quantitative indicators remain central to demonstrating empirical claims, while the interpretative depth of qualitative inquiry remains indispensable when social meanings, symbolic structures, or understudied phenomena require analytical attention. The expansion of digital communication further complicates this methodological landscape: contemporary media environments produce unprecedented volumes of unstructured textual data, making manual qualitative analysis impractical and reinforcing the need for hybrid strategies that integrate computational tools with interpretive reasoning.

CS&SE@SW 2025: 8th Workshop for Young Scientists in Computer Science & Software Engineering, November 21, 2025, Kryvyi Rih, Ukraine
https://doi.org/10.55056/cssesw2025/paper41
lytvynenko.oleksandr.15@gmail.com (O. Lytvynenko); miroslavakukhta@gmail.com (M. Kukhta); yeninmaksym@gmail.com (M. Yenin)
https://scholar.google.com/citations?user=ZC5apwgAAAAJ (O. Lytvynenko); https://sociology.kpi.ua/en/faculty-members/kukhta (M. Kukhta); https://sociology.kpi.ua/faculty-members/maksym-yenin (M. Yenin)
0009-0002-9165-4482 (O. Lytvynenko); 0000-0003-4663-9670 (M. Kukhta); 0000-0002-3835-2429 (M. Yenin)

Natural language processing (NLP) represents one such bridge between qualitative and quantitative approaches. By transforming textual material into analysable units while preserving semantic content, NLP techniques allow sociologists to examine large-scale discourse without abandoning qualitative logic. Among these techniques, sentiment analysis is particularly relevant for studying affective orientations embedded in news media, political communication and public debate. In practice, however, sentiment scores may vary substantially depending on the preprocessing strategy applied to the text. This problem is especially salient for morphologically rich languages such as Ukrainian, where inflectional complexity complicates normalisation and lexical matching.

Existing scholarship offers important but partial foundations for addressing this challenge. Classical and contemporary sociological works document the long-standing tension between qualitative and quantitative methodologies, emphasising the role of narrative materials and interpretive understanding in empirical research [1]. At the same time, methodological and technical manuals in NLP demonstrate how computational tools can be used to process large textual corpora, including through tokenisation, lemmatisation, stemming and lexicon-based sentiment analysis [2, 3]. Yet, despite this progress, there is limited empirical evidence on how specific preprocessing choices affect sentiment outcomes in sociological applications, particularly for Ukrainian-language media data.

The present study addresses this methodological gap by evaluating how two common preprocessing techniques – lemmatisation and stemming – affect sentiment analysis outcomes when applied to Ukrainian online news headlines. The central research question asks whether different preprocessing strategies produce divergent sentiment scores and what these differences imply for sociological interpretation of public discourse. Empirically, the study constructs two parallel corpora of 990 scraped news headlines, each aligned with a corresponding Ukrainian sentiment dictionary, and compares the resulting sentiment distributions obtained via a modified VADER implementation.

The contribution of the article is twofold. First, it provides replicable evidence that preprocessing decisions – specifically the choice between lemmatisation and stemming – can lead to measurable differences in sentiment classification in a morphologically rich language. Second, it offers a transparent workflow that integrates qualitative interpretive logic with computational efficiency, demonstrating how NLP tools may be systematically incorporated into sociological research on media discourse.

2. Literature review

Research on the relationship between qualitative and quantitative methodologies in sociology reflects a long-standing epistemological divide evident in both classical and contemporary scholarship. Foundational qualitative traditions emphasised subjective meaning and the interpretive understanding of social action. Weber's [4] methodological writings offered the earliest systematic argument for an interpretive approach, insisting that sociology must grasp the symbolic structures that orient behaviour rather than reduce them to numerical indicators [4]. This interpretive tradition was further institutionalised by early empirical work of Thomas and Znaniecki [5]. Their study, *The Polish Peasant in Europe and America*, demonstrated that personal documents, letters, and life stories serve as primary empirical materials through which sociologists can reconstruct social experiences and collective identities. As Cersosimo [6] notes, the life-story method explicitly situates subjective experience within sociological explanation. Schutz [7] further notes that each person's stream of consciousness is unique, structured by their own memories, perceptions, and acts of attention.

At the same time, contemporary methodological dictionaries reveal the endurance of debates regarding the appropriate balance between qualitative and quantitative approaches. Bruce and Yearley [1] observe that the preference for either strategy often reflects deeper theoretical commitments and that sociological schools continue to negotiate the boundaries between interpretive inquiry and statistical formalisation. This tension intensified in the early twenty-first century as political, economic, and technological actors increasingly demanded quantifiable evidence, contributing to the widespread adoption of computational techniques in empirical research.

Parallel to these disciplinary developments, the rapid expansion of digital communication funda-

mentally reshaped the methodological landscape. Programming books such as *Natural Language Processing with Python* [2] and *Practical NLP* [3], along with the Python ecosystem (BeautifulSoup, spaCy, NLTK, Pytesseract), have systematised computational approaches, providing researchers with tools for large-scale text processing, tokenisation, lemmatisation, stemming and feature extraction.

Despite the availability of these tools, several methodological gaps remain in the sociological literature. Existing studies rarely address the implications of preprocessing decisions – particularly morphological normalisation – for the validity of sentiment analysis. The majority of available sentiment-analysis methods and lexicons are optimised for English; tools for morphologically complex languages, such as Ukrainian, remain limited. Existing resources, including Dyomkin's [8] Ukrainian tone dictionary and Kupriienko's [9] stopword list, provide partial linguistic infrastructure but have not been systematically tested for sociological use.

Thus, although prior scholarship has addressed the epistemological tension between qualitative and quantitative paradigms [1, 4], the interpretive role of narrative materials [6, 5], and the emergence of NLP tools [2, 3], no existing work systematically evaluates how different preprocessing strategies alter sentiment analysis outcomes in Ukrainian-language sociological datasets. The present study, therefore, positions itself as a methodological contribution.

3. Conceptual background – NLP as a bridge between qualitative and quantitative methodologies

Contemporary sociology increasingly benefits from methodological strategies that integrate qualitative interpretation with quantitative analytical procedures. Classical qualitative research relied on the collection and understanding of language-based materials, treating text as a symbolic system through which actors construct and communicate meaning. In digital environments, innovative methodological combinations require not only interpreting text in its original form but also transforming linguistic elements into numeric representations suitable for algorithmic processing.

Content analysis illustrates this transformation particularly well. Historically rooted in document analysis and life-story materials, it depended heavily on the researcher's interpretive labour. With the expansion of digital communication and the emergence of Big Data, manual qualitative coding became increasingly inefficient. As a result, automated procedures – extraction, classification, scoring, and structured parsing – have been incorporated into content-analytic workflows, preserving interpretive depth while enabling systematic, replicable analysis of extensive datasets [1, 6].

Broader changes in communication practices further amplify the need for these hybrid approaches. As noted by Toffler [10], advanced societies increasingly centre on the production and consumption of experiences rather than on material goods alone. In digital environments, communication becomes a space where emotional engagement and symbolic meaning acquire strategic value. In this context, sentiment analysis is a relevant sociological tool for identifying emotional orientations embedded in public discourse.

Despite the computational procedures involved, sentiment analysis in this study remains aligned with qualitative reasoning. Words continue to function as the primary units of meaning, even when embedded in lexicons or processed through NLP pipelines. The preprocessing steps – cleaning, tokenisation, lemmatisation and stemming – aim to preserve semantic context, reduce noise and ensure interpretive coherence. In this sense, NLP serves as a methodological bridge, enabling the extension of qualitative inquiry into large-scale empirical settings. Preprocessing is consistent with best practices in NLP, where it is considered a crucial factor in shaping downstream analytical outcomes [2, 3].

4. Processing

4.1. Technical environment

All computational procedures in this study were carried out in Python using interpreter version 3.12.6 [11]. Python was selected due to its widespread adoption in data science and text analytics, its extensive ecosystem of specialised libraries and its ability to integrate data collection, preprocessing, modelling and visualisation within a single workflow.

The implementation relies on several key libraries. BeautifulSoup4 [12], together with standard networking modules including urllib [13], was used to parse HTML documents and retrieve web pages. OCR was performed using the Tesseract engine [14, 15] accessed via the Pytesseract wrapper [16], supplemented by the Pillow library [17] for image handling. Where necessary, Selenium [18] was considered as a complement for platforms restricting standard HTTP-based scraping.

4.2. Reproducibility

To ensure the reproducibility of the computational workflow, all scripts were executed in a controlled Python 3.12.6 environment with fixed versions of all dependencies. The project repository [19] contains the complete source code, configuration files and example datasets required to replicate each stage of the pipeline. This reproducibility protocol adheres to open-science standards, facilitating transparent and independent verification of the results.

4.3. Web data collection (scraping)

Given the focus on online news discourse, the empirical material for this study consists of headlines and associated metadata retrieved from publicly accessible news websites. Data collection followed a structured scraping strategy in which web pages were requested programmatically and their HTML content parsed using BeautifulSoup4 [12]. Illustrative examples, including code adapted from Vajjala et al. [3], demonstrate how the parser targets particular `<div>` blocks, extracts headline text, and collects supplementary metadata such as hyperlinks and timestamps (see listing 1).

Listing 1: Scraping using BeautifulSoup4 (illustrative example)

```
from bs4 import BeautifulSoup
from urllib.request import urlopen

html = urlopen("https://example.com").read()
soup = BeautifulSoup(html, "html.parser")

for article in soup.find_all("div", class_=lambda c: c in CARD_CLASSES):
    headline = article.find("span", class_=lambda c: c in HEADLINE_CLASSES)
    if headline:
        print(headline.get_text().strip())
```

To enhance robustness, the extraction logic applied predicate functions (e.g., lambda expressions) to identify container elements whose class attributes partially matched predefined patterns, enabling the scraper to tolerate minor layout changes. Only websites that permitted automated access were included, and scraping frequency was limited in line with responsible data-collection practices.

4.4. Ethical note

All scraping activities were conducted in accordance with the publicly available terms of use of the selected news platforms. No personal or user-generated data were collected, stored, or processed at any stage. The dataset consists exclusively of publicly available headlines and metadata, used solely for academic and methodological purposes.

4.5. Extraction of text from images

Not all relevant information in online media appears as plain HTML text. News items, opinion pieces, and thematic reports frequently include image-based materials that contain textual elements relevant to sentiment and framing analysis. To avoid omitting such material, the study incorporated an OCR component. The Tesseract OCR engine [14, 15] was employed via the Pytesseract interface [16]. Images were opened and preprocessed using the Pillow library [17], and OCR was applied with Ukrainian language configuration.

4.6. OCR limitations

While OCR enhances dataset completeness, it also introduces potential distortions. Recognition accuracy varies depending on image quality, font characteristics, background contrast, and the presence of artefacts. Tesseract may misidentify characters, merge or split tokens incorrectly, or fail to detect diacritics. For this reason, OCR-derived strings were not employed in the actual analysis; they are presented only as an illustrative example of potential application.

4.7. Qualitative logic, numeric transformation and the function of NLP

In contemporary sociological research, qualitative and quantitative approaches should be seen as complementary rather than mutually exclusive. Content analysis provides a clear example of methodological transformation: algorithmic procedures have been integrated into content-analytic strategies, preserving interpretive depth while enabling systematic analysis [1, 6].

The removal of stopwords is a central preprocessing operation. As the NLTK library does not provide a built-in Ukrainian stopword list [20], an openly available Ukrainian stopword list [9] was incorporated and adapted to the needs of this study.

Lemmatisation was implemented using the transformer-based Ukrainian model in spaCy [21, 22]. The `uk_core_news_trf` pipeline performs tokenisation, part-of-speech tagging and lemma assignment, reducing inflected forms to their dictionary base. The Ukrainian tone dictionary [8] undergoes the same lemmatisation procedure. Aligning both the corpus and the tone dictionary at the lemma level minimises mismatches caused by inflection and supports more stable sentiment scoring.

Listing 2: spaCy Ukrainian transformer pipeline – lemmatisation (illustrative example)

```
import spacy

nlp = spacy.load("uk_core_news_trf")

for _, row in tonedict.iterrows():
    doc = nlp(row["word"])
    lemmatonedict[doc[0].lemma_] = float(row["score"])

for article in articles:
    doc = nlp(article["headline_clean"])
    article["headline_lemma"] = " ".join(tk.lemma_ for tk in doc)
```

The second normalisation strategy is stemming. In contrast to lemmatisation, stemming reduces tokens to shorter, root-like forms without referring to a lexical database [23]. As no Ukrainian-specific stemmer is currently available within the NLTK library [20], the pipeline employs the nearest available Snowball stemmer configuration. This choice is motivated solely by shared morphological properties within the East Slavic language group and represents a technical approximation; its limitations are explicitly acknowledged in Section 9. Data manipulation was supported by the pandas library [24].

Listing 3: NLTK-based stemming pipeline (illustrative example)

```
from nltk.stem.snowball import SnowballStemmer
```

```
stemmer = SnowballStemmer("russian")  # Ukrainian not available

for _, row in tonedict.iterrows():
    stemtonedict[stemmer.stem(row["word"])] = float(row["score"])

for article in articles:
    words = article["headline_lemma"].split()
    article["headline_stem"] = " ".join(stemmer.stem(w) for w in words)
```

As a result of this pipeline, the study produces two parallel representations of the same corpus. The first combines lemmatised headlines with a lemmatised tone dictionary, while the second pairs stemmed headlines with a stem-based tone dictionary. Since the underlying set of articles and headlines remains identical, any differences in sentiment scores can be attributed specifically to the choice between lemmatisation and stemming.

4.8. Text analysis

A central instrument in the empirical analysis is VADER [25]. Although distributed through the NLTK library and widely used in sentiment analysis, VADER is a lexicon-based system assigning affective weights to predefined lexical items. In this study, the lexicon was adapted by integrating lemmatised and stemmed forms from the Ukrainian tone dictionary [8]. Transformer-based architectures – while not used in the empirical analysis – provide conceptual background for understanding the transformer-based spaCy pipeline applied in Ukrainian lemmatisation [21, 22].

4.9. Sentiment analysis

The sentiment analysis is conducted using VADER [25], integrated through the NLTK library [20]. Two customised lexicons – derived from Dyomkin's [8] Ukrainian tone dictionary – were prepared: a lemmatised version aligned with the lemmatised corpus and a stem-based version aligned with the stemmed corpus. Loading each lexicon into a separate VADER analyser ensures consistency between the morphological form of tokens and the lexical entries available for matching.

Listing 4: Sentiment analysis using VADER with custom Ukrainian lexicons (illustrative example)

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer

SIA_lemma = SentimentIntensityAnalyzer()
SIA_lemma.lexicon.update(lemma_tonedict)

SIA_stem = SentimentIntensityAnalyzer()
SIA_stem.lexicon.update(stem_tonedict)

for article in articles:
    lps = SIA_lemma.polarity_scores(article["headline_lemma"])
    sps = SIA_stem.polarity_scores(article["headline_stem"])
    article["lemma_compound"] = lps["compound"]
    article["stem_compound"]  = sps["compound"]
```

The script iterates through all headlines and computes the four standard VADER metrics – negative, neutral, positive and compound sentiment scores – for each morphological variant of each headline. The intermediate results indicate that 107 out of 990 headlines (10.8%) receive different compound sentiment values depending solely on the choice between lemmatisation and stemming. This proportion demonstrates that morphological normalisation is not a neutral preprocessing step but a factor capable of altering sentiment classifications.

5. Visualisation

The final stage of the study involves the development of an interactive interface that enables researchers to examine the processed dataset and compare sentiment outputs generated through different preprocessing strategies. A lightweight dashboard was implemented using the Shiny framework for Python [26, 27]. The dashboard allows users to browse the dataset by selecting article identifiers, conducting keyword searches or navigating through the complete list of processed items. For each selected headline, the interface displays the original text, the publication timestamp and both sentiment-analysis outcomes, with divergences highlighted automatically. The dashboard is converted into a browser-executable format through the Shinylive package [27] and is accessible at [28]. Visualisations of the results were created using Tableau [29].Figures 1 and 2 present illustrative plots used in the analysis.

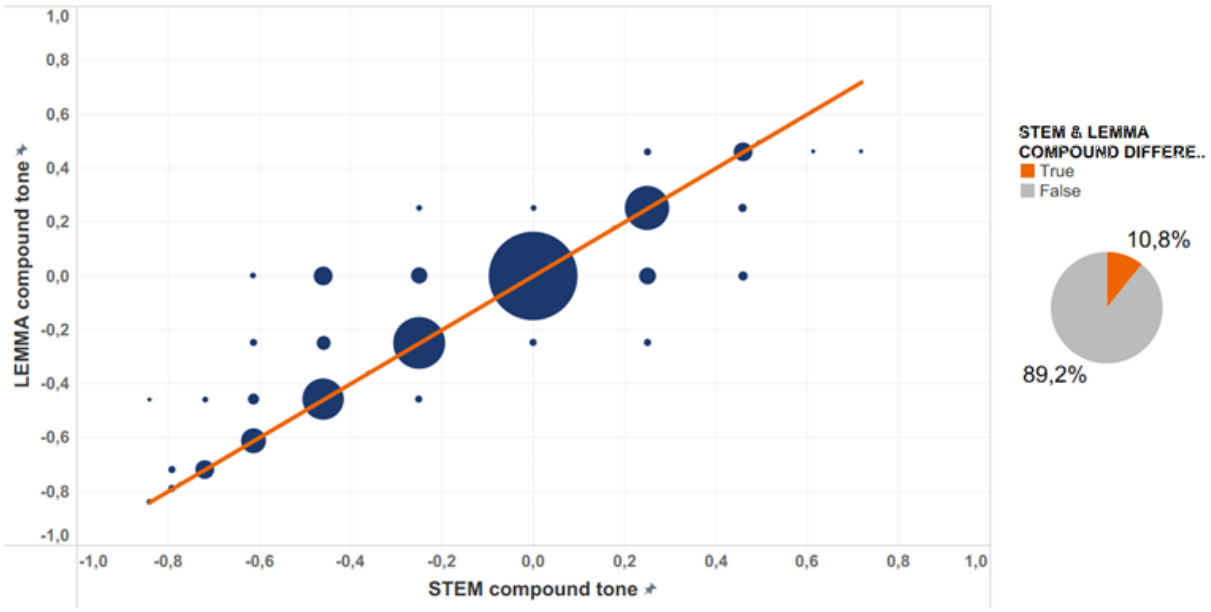

Figure 1: Scatter plot of compound sentiment scores with reference line.

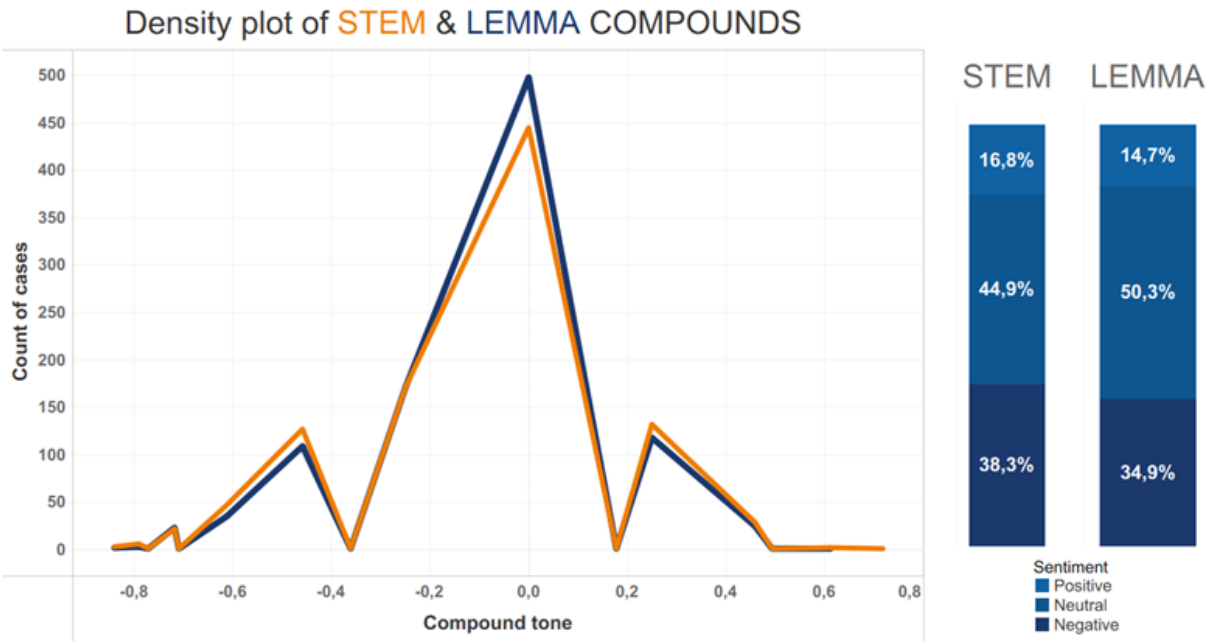

Figure 2: Density plot of lemma- and stem-based compound sentiment scores.

6. Formalisation of results

The comparison of lemmatisation and stemming demonstrates a measurable divergence in sentiment classification outcomes. Among the 990 processed headlines, 107 (10.8%) produced different compound sentiment scores when analysed using the lemmatised versus the stemmed version of the text. This proportion shows that morphological normalisation is not a neutral or purely technical decision but a factor capable of shaping analytical results. The sensitivity of sentiment classification to preprocessing choices is particularly relevant in studies examining polarity shifts, temporal dynamics or topic-specific affective patterns.

The findings therefore indicate that researchers working with morphologically rich languages must explicitly justify their preprocessing strategies and remain attentive to their analytical consequences. Since lexicon-based sentiment scoring depends directly on lexical matching, even small variations in token form can alter sentiment outcomes.

7. Discussion

The results of the study confirm that preprocessing is a substantive methodological decision rather than a neutral technical operation. The divergence of 107 compound sentiment scores out of 990 analysed headlines demonstrates that morphological normalisation directly shapes analytical outcomes when working with Ukrainian-language text. This finding aligns with existing methodological literature emphasising that text preprocessing has interpretive consequences within NLP workflows [2, 3].

From a sociological perspective, this sensitivity is particularly important when sentiment analysis is used to study public discourse, political communication, crisis framing or media polarisation. If sentiment classifications shift due to preprocessing alone, then downstream interpretations of emotional tone, public reactions or topic salience may also differ. Consequently, sentiment outputs cannot be treated as inherently objective indicators unless the preprocessing pipeline is explicitly documented and justified.

The findings further illustrate that computational methods can reinforce rather than replace qualitative logic. Lemmatisation, in particular, retains semantic consistency and therefore aligns with the interpretive requirements of sociological analysis. Stemming, while computationally efficient, may oversimplify linguistic structure and distort lexical meaning in morphologically rich languages. Overall, the results encourage a shift in sociological NLP practice toward methodological reflexivity: researchers must treat preprocessing choices as theoretical decisions that influence empirical interpretation.

8. Practical implications

The study provides several practical contributions for computational sociology and digital media research. First, it offers a reproducible pipeline for collecting, cleaning and processing Ukrainian-language news content using openly available tools, thereby lowering the technical threshold for scholars entering NLP-based research. Second, the empirical comparison of preprocessing strategies shows when lemmatisation should be preferred – particularly in studies requiring semantic precision, interpretive depth or sensitivity to contextual nuance. Third, higher-speed workflows may still employ stemming, provided that its limitations are acknowledged and results interpreted cautiously.

The adaptation of VADER through custom sentiment dictionaries illustrates how existing tools can be localised for Ukrainian without requiring access to large labelled corpora or high-performance computational resources. More broadly, the workflow presented in this study may serve as a template for integrating NLP methods into qualitative–computational research frameworks in sociology, communication studies and political science.

9. Limitations

Several methodological limitations should be acknowledged. First, the sentiment dictionary used as the basis for scoring is illustrative and has not undergone large-scale empirical validation. Second, the dataset includes a finite selection of 990 headlines from a limited number of online news platforms. Headlines provide only a condensed representation of larger textual narratives; the scope of the analysis does not extend to full articles or multimodal forms of communication.

Third, stemming relied on a non-Ukrainian morphological algorithm due to the absence of a publicly available Ukrainian stemmer, which may introduce linguistic distortions. Fourth, OCR recognition accuracy varies according to image quality and typography; OCR-derived errors may propagate into subsequent stages and their potential impact should be considered when evaluating the outcomes.

These limitations do not undermine the methodological contribution of the study. Rather, they indicate that the project functions as a test implementation aimed at demonstrating how preprocessing choices influence sentiment-analysis results in Ukrainian-language data.

10. Conclusions

This study demonstrates how Natural Language Processing techniques can expand the analytical capacity of qualitative sociological research by enabling systematic, large-scale examination of textual data. Using online news headlines as empirical material, the analysis compared two preprocessing strategies – lemmatisation and stemming – to assess how different forms of morphological normalisation influence sentiment-analysis outcomes. The results indicate that out of 990 headlines, 107 (10.8%) generated divergent compound sentiment scores depending solely on whether lemmatisation or stemming was applied. This outcome highlights the methodological significance of preprocessing choices and underscores the need for transparency when incorporating computational tools into sociological research designs.

The study's methodological contribution lies in showing how qualitative logic – centred on meaning, semantic nuance and interpretive context – can be preserved within computational pipelines. By applying lemmatisation and stemming to both the corpus and the sentiment dictionary, the article demonstrates that algorithmic procedures reshape, rather than replace, qualitative interpretation. This affirms the value of NLP as a bridge connecting interpretive and quantitative approaches.

These limitations suggest directions for future research. Larger datasets, full-text corpora and advanced contextual models – such as transformer-based architectures adapted for Ukrainian – may yield more nuanced insight into affective patterns in media communication. Further work is also needed to create and validate sentiment lexicons tailored to Ukrainian morphology, genre-specific vocabulary and domain-dependent usage.

Despite being a test implementation, the project illustrates how sociologists can incorporate computational tools into qualitative research in a transparent, replicable and methodologically grounded manner. The results highlight the importance of methodological reflexivity when applying NLP techniques and demonstrate that hybrid qualitative–computational approaches can enhance research on digital media, public discourse and communication dynamics – provided that preprocessing pipelines are explicitly documented and critically evaluated. The sentiment dictionary used in this project serves only as an example and requires substantial refinement; the entire computational workflow should be interpreted as a demonstrational prototype.

Author contributions

The following statements should be used Conceptualization, Oleksandr Lytvynenko and Myroslava Kukhta; methodology, Oleksandr Lytvynenko; software, Oleksandr Lytvynenko; validation, Myroslava Kukhta and Maksym Yenin; formal analysis, Oleksandr Lytvynenko; investigation, Oleksandr Lytvynenko; resources, Myroslava Kukhta; data curation, Oleksandr Lytvynenko; writing—original draft preparation, Oleksandr Lytvynenko; writing—review and editing, Oleksandr Lytvynenko and Myroslava

Kukhta; visualization, Oleksandr Lytvynenko; supervision, Myroslava Kukhta and Maksym Yenin; project administration, Maksym Yenin.

Funding

This work was supported by the R&D project No. 2916f of the National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute" – "The role of artificial intelligence in shaping educational policies for the development and preservation of human capital" (No. DR 0126U001611), and by the project "Pan-European Network of a Multisectoral Master's Programme in Responsible Artificial Intelligence (PANORAIMA)" (Grant No. 101189994).

Data availability statement

The project repository [19] contains the complete source code, configuration files and example datasets required to replicate each stage of the pipeline. Restrictions apply to the availability of the full news-headline dataset due to third-party platform terms of use.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgments

The authors thank the developers of the open-source tools used in this study.

Declaration on Generative AI

The authors used generative AI tools only for language polishing of selected passages; all analytical, methodological and interpretive work is the authors' own.

References

[1] S. Bruce, S. Yearley, The SAGE Dictionary of Sociology, SAGE Publications Ltd, London, 2006. doi:10.4135/9781446279137.

[2] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, O'Reilly Media, Sebastopol, 2009.

[3] S. Vajjala, B. Majumder, A. Gupta, H. Surana, Practical Natural Language Processing, O'Reilly Media, Sebastopol, 2020.

[4] M. Weber, Economy and Society, University of California Press, Berkeley, Los Angeles, London, 1978. URL: https://archive.org/details/MaxWeberEconomyAndSociety.

[5] W. I. Thomas, F. Znaniecki, The Polish peasant in Europe and America; monograph of an immigrant group, Richard G. Badger, Boston, 1918-20. URL: https://archive.org/details/polishpeasantine01thomuoft/polishpeasantine01thomuoft/.

[6] G. Cersosimo, The Polish Peasant in Europe and America. Some Remarks after One Hundred Years, Italian Sociological Review 10 (2020) 329–340. doi:10.13136/isr.v10i2s.347.

[7] A. Schutz, The Phenomenology of the Social World, Studies in Phenomenology and Existential Philosophy, Northwestern University Press, Evanston, 1967.

[8] V. Dyomkin, lang-uk/tone-dict-uk: Ukrainian tone dictionary, 2016. URL: https://github.com/lang-uk/tone-dict-uk.

[9] S. Kupriienko, skupriienko/Ukrainian-Stopwords: the list of ~2000 ukrainian stopwords (with numbers), 2021. URL: https://github.com/skupriienko/Ukrainian-Stopwords.
[10] A. Toffler, Future Shock, Bantam Book, Toronto, New York, London, 1970. URL: https://archive.org/details/FutureShock-Toffler.
[11] Python Software Foundation, Welcome to Python.org, 2026. URL: https://www.python.org.
[12] L. Richardson, beautifulsoup4 · PyPI, 2026. URL: https://pypi.org/project/beautifulsoup4/.
[13] Python Software Foundation, urllib.request — Extensible library for opening URLs — Python 3.14.7 documentation, 2026. URL: https://docs.python.org/3/library/urllib.request.html.
[14] R. W. Smith, Hybrid Page Layout Analysis via Tab-Stop Detection, in: 2009 10th International Conference on Document Analysis and Recognition, 2009, pp. 241–245. doi:10.1109/ICDAR.2009.257.
[15] R. Smith, D. Antonova, D.-S. Lee, Adapting the Tesseract open source OCR engine for multilingual OCR, in: Proceedings of the International Workshop on Multilingual OCR, MOCR '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 1. doi:10.1145/1577802.1577804.
[16] S. Hoffstaetter, madmaze/pytesseract: A Python wrapper for Google Tesseract, 2025. URL: https://github.com/madmaze/pytesseract.
[17] J. A. Clark, contributors, Python pillow, 2028. URL: https://python-pillow.org.
[18] Software Freedom Conservancy, Selenium, 2026. URL: https://www.selenium.dev.
[19] O. Lytvynenko, Bigtoothdev/nlp-sociology-methodology, 2026. URL: https://github.com/BigToothDev/nlp-sociology-methodologyy.
[20] NLTK Project, Nltk :: Natural language toolkit, 2025. URL: https://www.nltk.org.
[21] Explosion, spaCy · Industrial-strength Natural Language Processing in Python, 2026. URL: https://spacy.io.
[22] Explosion, Ukrainian · spaCy Models Documentation, 2026. URL: https://spacy.io/models/uk.
[23] Snowball, 2001. URL: https://snowballstem.org/.
[24] The Pandas Development Team, pandas-dev/pandas: Pandas (Version v3.0.5) [Computer software], Zenodo, 2026. doi:10.5281/zenodo.3509134.
[25] C. Hutto, E. Gilbert, VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text, in: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM), volume 8(1), 2014, pp. 216–225. doi:10.1609/icwsm.v8i1.14550.
[26] Posit Software, Shiny for Python, 2024. URL: https://shiny.posit.co/py/.
[27] B. Schloerke, W. Chang, G. Stagg, G. Aden-Buie, shinylive: Run 'shiny' Applications in the Browser, 2026. URL: https://posit-dev.github.io/r-shinylive/, r package version 0.5.0.
[28] O. Lytvynenko, DataExplorer, 2026. URL: https://bigtoothdev.github.io/nlp-sociology-methodology/y.
[29] Salesforce, Inc., Business Intelligence and Analytics Software | Tableau, 2026. URL: https://www.tableau.com.