Topic modelling of Ukrainian folk songs: A case study on Podillia region

Olha B. Petrovych[1,2]

[1]*Estonian Literary Museum, 42 Vanemuise Str., Tartu, 51003, Estonia*
[2]*Vinnytsia Mykhailo Kotsiubynskyi State Pedagogical University, 32 Ostrozhskogo Str., Vinnytsia, 21100, Ukraine*

Abstract

This study explores the thematic structures and motifs encoded in the folk songs of the Podillia region, utilizing computational methods to uncover patterns of cultural memory and narrative tradition. The primary research questions addressed are: (1) What themes and motifs emerge from the analysis of Podillia folk songs using topic modelling with Latent Dirichlet Allocation (LDA)? (2) How do these computational findings align or diverge from traditional folkloristic classifications? (3) What challenges arise in adapting computational methods to Ukrainian folk song corpora, and how can they be addressed? The dataset, comprising 2762 folk songs, was subjected to preprocessing steps including tokenization, lemmatization, and stopword removal. A document-term matrix was constructed and filtered to focus on meaningful terms. After applying LDA, the top 20 latent topics were filtered, each characterized by distinct thematic clusters and keywords. Results highlight recurrent themes such as seasonal cycles, family relationships, and social rituals. This research not only provides new insights into the thematic richness of Podillia folk songs but also demonstrates the potential of computational folkloristics to complement traditional methodologies. However, limitations such as data sparsity, Ukrainian language-specific challenges, and the interpretability of computational outputs are noted. Future research should focus on refining methodologies and integrating hybrid approaches to deepen our understanding of cultural narratives encoded in folk traditions.

Keywords

topic modelling, Latent Dirichlet Allocation (LDA), folk songs, Podillia region

1. Introduction

Folk songs are a vital component of intangible cultural heritage, serving as repositories of historical memory, collective emotions, and shared values within communities. In Ukraine's Podillia region, these songs preserve local traditions, moral codes, and social practices. This oral tradition contributes to broader discussions about the role of folk culture in shaping national identity [1, 2, 3] and collective memory [4]. However, the vastness and thematic diversity of this oral tradition stimulate researchers to systematically categorize and interpret the main motifs and narratives.

Recent advances in computational methods provide tools for addressing these challenges. Among these methods, topic modelling has gained prominence. This method is designed to uncover latent themes within large unstructured textual corpora. LDA (Latent Dirichlet Allocation), one of the most widely used topic modelling techniques, operates by identifying patterns in word co-occurrence to infer topics as clusters of related terms. Blei [5] highlights that topic modelling can "discover the main themes that pervade a large and otherwise unstructured collection of documents", making it a promising approach for analysing large folk song corpus. Studies by Murakami et al. [6] and DiMaggio et al. [7] further demonstrate the ability of LDA to identify interpretable themes across diverse datasets, bridging qualitative and quantitative approaches to textual analysis.

The importance of this research lies in its potential to enhance our understanding of Podillia's cultural heritage through the application of computational methods in folklore studies. However, as highlighted by Petrovych et al. [8], there are linguistic challenges that may not align well with the capabilities of

CS&SE@SW 2024: 7th Workshop for Young Scientists in Computer Science & Software Engineering, December 27, 2024, Kryvyi Rih, Ukraine
olha.petrovych@vspu.edu.ua (O. B. Petrovych)
https://sites.google.com/vspu.edu.ua/kafukrlit/first/petrovich (O. B. Petrovych)
0000-0002-7185-3823 (O. B. Petrovych)

modern natural language processing (NLP) tools. Folk songs often employ colloquial language and regional dialects, which can lead to inaccuracies in processes such as lemmatization and stopword removal. Furthermore, the use of metaphorical and symbolic language presents additional difficulties, as algorithms may struggle to interpret these elements without adequate contextual knowledge.

This study aims to explore the thematic structures embedded in Podillia region folk songs by applying LDA-based topic modelling. Specifically, the research questions guiding this study are:

1. What themes and motifs emerge from the analysis of Podillia folk songs using topic modelling with LDA?
2. How do these computational findings align or diverge from traditional folkloristic classifications?
3. What challenges arise in adapting computational methods to Ukrainian folk song corpora, and how can they be addressed?

By addressing these challenges, the research aim and questions, this study explores the thematic diversity of Podillia folk songs and the methodological peculiarities of analysing cultural heritage through digital means.

2. Topic modelling in literature, music, and folklore research

Topic modelling, particularly LDA, has emerged as a powerful method for uncovering latent semantic structures within large text corpora. Egger [9] provides a broader theoretical framework for topic modelling, emphasizing its ability to uncover hidden semantic structures and its adaptability across diverse datasets. Sobchuk and Šeļa [10] explored how LDA and similar models could identify subtle thematic nuances within large corpora, revealing patterns often overlooked in traditional literary analysis. Similarly, Nylander and Holmer [11] applied topic modelling with LDA to analyse educational materials, focusing on folk high school cultural and educational texts to uncover thematic structures.

Topic modelling has also been employed to analyse sentiment and emotions in song lyrics. Devi and Saharia [12] utilized LDA to classify sentiments in songs, demonstrating the method's ability to handle abstract and emotional content. Similarly, Dakshina and Sridhar [13] investigated emotion recognition in lyrics using LDA, demonstrating its efficacy in identifying complex emotional patterns.

Laoh et al. [14] conducted a study on Indonesian song lyrics, employing LDA to explore their thematic content. Their findings underscore the role of computational models in examining cultural narratives across regional contexts. Wanjantuk et al. [15] extended this approach to Thai songs, unveiling themes and emotions through topic modelling, further validating the method's cross-cultural applicability.

Panda et al. [16] demonstrated the integration of topic modelling with visualization techniques to explore music genres, providing insights into the underlying thematic structure of musical texts. This combination of analysis and visualization offers a compelling way to interpret the diversity of cultural content.

The application of topic modelling in folklore studies has proven fruitful. Sarv's [17] work on Estonian runosongs demonstrated the potential of LDA in identifying thematic patterns in oral traditions. Similarly, Strle and Marolt [18] highlighted the role of computational methods in semantic analysis within folkloristics, emphasizing their utility in bridging qualitative and quantitative approaches. These studies provide a roadmap for the application of LDA to similar cultural datasets and underline the importance of adapting methodologies to the linguistic and cultural specificities of the data.

While topic modelling has been widely adopted in various fields, its application to Ukrainian texts has been limited, often focusing on contemporary datasets. For instance, Kryndach et al. [19] explored semantic relationships in Ukrainian text using Word2Vec and machine learning, showcasing the adaptability of computational tools in handling Ukrainian linguistic nuances. Similarly, Khairova et al. [20] utilized LDA with Collapsed Gibbs Sampling to analyse war-related news, highlighting the method's capacity to uncover dominant themes in modern contexts. However, these studies primarily address current events, leaving the exploration of historical and cultural texts underdeveloped.

One of the key challenges in working with Ukrainian texts lies in the under-resourced nature of the language within NLP. Researchers often translate Ukrainian corpora into English to leverage pre-trained

models, as noted in Verbytska's [21] study on war narratives. This translation step, while practical, introduces potential biases and diminishes the authenticity of cultural context.

By identifying co-occurrence patterns of words, LDA can reveal underlying themes, making it a valuable tool for analysing cultural texts like folk songs. Despite its potential, there remains a significant gap in applying topic modelling to Ukrainian-language data, particularly within the domain of folklore. My study aims to address these limitations by applying LDA directly to a corpus of Ukrainian folk songs, preserving their linguistic and cultural integrity.

3. Data and methods

To analyse the topic modelling of the Ukrainian folk songs corpus, the R programming language (version 4.4.1) along with RStudio (version: 2024.04.2+764) is used. The text corpus and R scripts to analyse the topic modelling of Podillia region folk songs are available at Zenodo [22].

This study employs a range of text analysis techniques, including text preprocessing, stopwords removal, tokenization, lemmatization, part-of-speech (PoS) tagging, and document-term matrix (DTM) construction. Topic modelling was performed using LDA with coherence evaluation metrics. Additionally, word embeddings were generated with the GloVe algorithm, and K-means clustering was applied for semantic grouping. Data visualization techniques, including bar charts, word clouds, principal component analysis (PCA), and heatmaps, were used to interpret and present findings. Below is a detailed account of the research methodology.

3.1. Corpus characteristics

This research is based on Ukrainian folk songs from the Podillia region collections [23, 24, 25]. The text corpus comprises 2762 songs containing a total of 52004 lines and 209075 tokens. The corpus of Podillia folk songs is in Ukrainian.

The corpus consists of folk songs from the Podillia region, located in central-western Ukraine, and is known for its rich folklore tradition. These songs feature colloquial language, dialectal expressions, and a strong reliance on symbolism. Such characteristics necessitate tailored preprocessing and advanced computational techniques to uncover meaningful patterns.

3.2. Preprocessing

Preprocessing is crucial for transforming raw textual data into a format suitable for computational analysis, particularly for folk song texts. There were such steps to prepare it for topic modelling:

1. *Stopword removal.* Stopwords were removed using both standard lists Stopwords-ISO from an external URL [26] and a manually curated set which was developed by close reading of text corpus to enhance filtering for the corpus. Stopword removal reduces noise in text analysis, a step underscored by Blei et al. [27] in their original work on LDA.
2. *Tokenization and kemmatization.* The corpus was tokenized into individual words and converted to lowercase. Lemmatization was applied using the Ukrainian UDPipe model [28]. Lemmatization ensures that morphological variations of words are unified into their base forms (e.g., *spivaie* 'sings' and *spivala* 'sang' to *spivaty* 'to sing'). PoS filtering was applied to retain content-rich words (nouns, adjectives, verbs, and adverbs), which contribute most to thematic content. Once the text has been preprocessed, different techniques can be used to represent the text as numerical features. One of them is DTM. A DTM was constructed where rows represent songs (documents) and columns represent unique terms (words). The matrix values corresponded to word frequencies in each document. Terms appearing in fewer than 10 percent of documents or with low absolute frequency (<100 occurrences) were filtered out to reduce noise and computational overhead. This step reduces sparsity and enhances topic coherence, as suggested by Steyvers and Griffiths [29].

3.3. Topic modelling with LDA

LDA [27], a generative probabilistic model, was employed to uncover latent themes (topics) across the songs in the corpus. It models each document as a mixture of topics and each topic as a distribution over words. Key parameters include:

1. *Coherence evaluation:* To evaluate the number of topics, coherence metrics [30] were employed. Coherence is assessed using four widely recognized metrics: Griffiths2004 [31], CaoJuan2009 [32], Arun2010 [33], and Deveaud2014 [34].
   Gibbs sampling method was used for inference, with a seed for reproducibility. The number of topics was optimized to 20 based. These metrics ensure that the identified topics are meaningful and distinct.
2. *Word embeddings and clustering*
   - Word embeddings were generated using the GloVe algorithm [35] to capture semantic relationships among words. These embeddings were trained on a term-co-occurrence matrix (TCM).
   - K-means clustering grouped words into semantically coherent clusters, aiding the interpretation of topics.

   Each topic was characterized by its top 20 terms (β values), and visualizations such as bar charts, PCA, and heatmaps of top terms [36] highlighted the keywords defining each topic.

4. Results and discussion

To uncover the thematic structure of Podillia region folk songs, LDA was applied to identify the top 20 topics within the corpus. Each topic is characterized by a set of keywords that represent recurring motifs and cultural themes embedded in the songs. The results offer a comprehensive overview of the semantic patterns present in the corpus, as visualized in figure 1.

The bar chart in figure 1 displays beta values on the x-axis, representing the relative importance of each keyword within its topic, while the y-axis lists the keywords. Each facet corresponds to a separate topic, arranged across 4 rows and 5 columns.

The filtered top 20 distinct thematic groups of Podillia region folk songs reflect key cultural and narrative elements, shedding light on the values, challenges, and sentiments expressed in traditional Ukrainian folklore. Each keyword list is associated with a folk motif or theme that best captures the essence of the topic. The central motifs, narrative patterns, and folkloristic significance of these themes are discussed below.

Topic 1: *zhinka* ‘woman’, *voda* ‘water’, *dobryi* ‘good’, *dodomu* ‘homeward’, *bihty* ‘to run’, *byty* ‘to beat’, *lito* ‘summer’, *bida* ‘trouble’, *vechir* ‘evening’, *sino* ‘hay’, *yar* ‘ravine’, *kosyty* ‘to mow’, *prynesty* ‘to bring’, *mynaty* ‘to pass’, *myty* ‘to wash’, *zolotyi* ‘golden’, *liahty* ‘to lie down’, *zillia* ‘herbs’, *liudy* ‘people’, *litaty* ‘to fly’.

This theme centres on a pastoral and domestic context, highlighting the lives of women in rural settings. The central motifs reflect daily struggles and the cyclicality of village life. Elements such as ‘to bring’, ‘to wash’, and ‘to lie down’ signify domestic labour and routines, while ‘summer’, ‘to mow’, and ‘hay’ evoke seasonal agricultural work tied to the rhythms of nature. Transitions like ‘to pass’, ‘to run’, and ‘homeward’ suggest movement, whether literal or symbolic. The dualities resonate with the complexities of rural life, where natural elements serve both as sustenance and as tools in human conflicts. For example, water often connects with purification in rituals. Herbs, frequently associated with healing, may also imply negative connotations, such as their use in poisons. The motifs of ‘trouble’, ‘to beat’, and ‘to lie down’ point to themes of domestic abuse and violation. Meanwhile, the mention of ‘people’ – often appearing alongside ‘good’ – reflects communal dynamics in village life. This pairing suggests an appeal to the community for support, yet it also highlights the passive, gossipy nature of rural society. The topic emphasizes the tension between community and individuality. This theme resonates with motifs of the pervasive hardships faced by women in traditional societies.

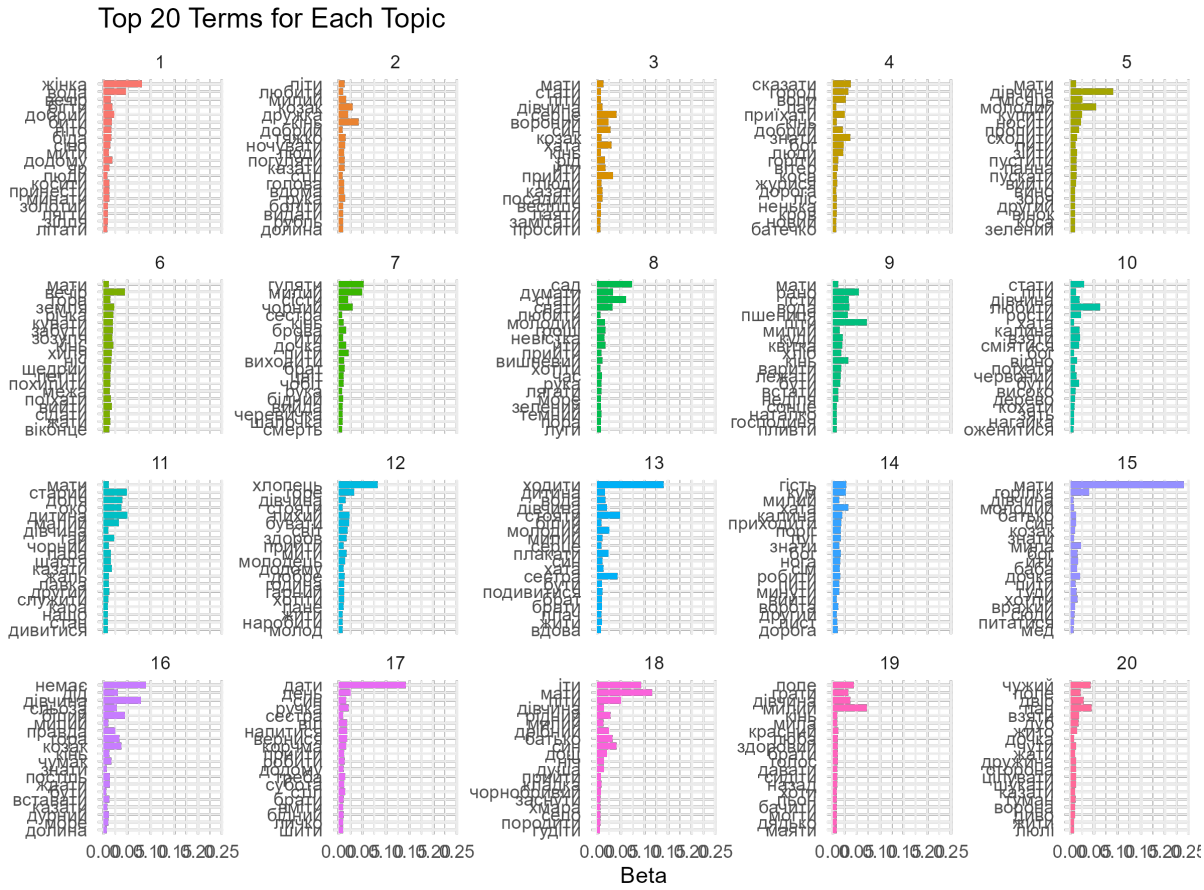

Figure 1: The distribution of the top 20 distinct topics with 20 keywords in each identified in Podillia region folk songs using LDA. The x-axis represents beta values. Each panel visualizes one topic, highlighting the key terms contributing to its thematic structure.

Topic 2: *kin* 'horse', *kozak* 'Cossack', *druzhka* 'bridesmaid', *mylyi* 'beloved', *tiazhko* 'hard', *ruka* 'hand', *nochuvaty* 'to spend the night', *pohuliaty* 'to revel', *pity* 'to go', *kazaty* 'to say', *holova* 'head', *vdoma* 'at home', *liudy* 'people', *bolity* 'to ache', *vydaty* 'to give', *rubl* 'ruble', *dolyna* 'valley', *liubyty* 'to love', *stil* 'table', *dobryi* 'good'.

This theme centres on the Cossack figure and his romantic interactions. The keywords suggest a narrative pattern that begins with courtship and intimacy ('to go', 'to say') and implies premarital sex ('to love', 'to spend the night'). It then follows with the betrayal of the girl, symbolized by the transactional act of the Cossack giving her money for the night ('to give', 'ruble'). The Cossack, astride his horse, departs, leaving behind the girl who loves him. The emotional aftermath is captured in words like 'hard', 'head', and 'to ache', reflecting the girl's worries, hardships, and emotional scars from abandonment. The Cossack's moments of carefree indulgence, signified by 'to revel' and 'table', contrast with the emotional burden carried by his beloved. The horse is a vital symbol, representing the Cossack's mobility, independence, and readiness to journey onward, underscored by his movement through the 'valley'. This theme encapsulates the archetypal narrative of romantic betrayal in Ukrainian folklore, where the Cossack, embodying freedom and individuality, deserts his beloved to pursue his wandering existence.

Topic 3: *sertse* 'heart', *pryity* 'to come', *khata* 'home', *syn* 'son', *voronyi* 'raven', *ity* 'to go', *rid* 'kin', *maty* 'mother', *kazaty* 'to say', *posadyty* 'to plant', *divchyna* 'girl', *liudy* 'people', *kozak* 'Cossack', *vesillia* 'wedding', *staty* 'to become', *laiaty* 'to scold', *pity* 'to go', *kin* 'horse', *zamitaty* 'to sweep', *prosyty* 'to ask'.

This theme reflects the celebratory and ritualistic aspects of weddings, intertwining within familial and romantic relationships. The 'home' serves as the narrative's central axis, anchoring key event such

as 'weddings'. Words like 'to ask' and 'to become' directly connect with the rituals and transitions surrounding marriage. The dynamic between 'kin', 'mother', and 'son', along with new responsibilities like 'to sweep' and 'to plant', highlights the changing roles and expectations placed upon a new family member. This readiness, or lack thereof, can often be a source of familial tension, reflected in words like 'to say' and 'to scold'. These elements blend to form a narrative of communal identity and personal emotions, capturing the balance between collective traditions and individual challenges. Keywords like the 'raven', 'horse', and 'Cossack' evoke notions of freedom and a single, unbound life, standing in contrast to the ties of family and marital duties.

Topic 4: *skazaty* 'to say', *znaty* 'to know', *horod* 'field', *voly* 'oxen', *pryikhaty* 'to arrive', *boh* 'God', *liudy* 'people', *dobryi* 'good', *hority* 'to burn', *viter* 'wind', *zhurysia* 'to grieve', *kosa* 'braid', *kin* 'horse', *lis* 'forest', *nenka* 'mother', *krov* 'blood', *novyi* 'new', *doroha* 'road', *pan* 'lord', *batechko* 'father'.

The theme resonates with the motif of difficult family life away from parents (represented by 'mother' and 'father'). In these narratives, the husband often appropriates the wife's belongings, physically abuses her, or forces her into submission. The wife's 'braid', symbolizing her cherished beauty, is compared to a broom used for sweeping the house. The husband might compel the wife to set out on a journey ('new', 'road') to collect livestock ('oxen', 'horse') from her parents, ostensibly to improve their household's fortune and to appease him. However, upon her return ('to arrive') with the goods, the story takes a tragic turn: he kills her ('blood') during a new quarrel. The mention of 'good' 'people' portrays them as informants who deliver the news of the murder ('to say', 'to know') to the wife's grieving mother ('to grieve'), rather than as active agents of help or justice. The keyword 'oxen' also alludes to Chumak songs, suggesting narratives of trials and hardships faced by Chumaks (itinerant traders and salt carriers in Ukrainian tradition). If a Chumak dies on the 'road', it is usually said that he has given his spirit to God ('God'). This topic reflects a common folkloristic pattern where natural forces ('to burn', 'wind') and human destiny intertwine, often with omens of tragedy or loss. Stories of trials faced by individuals or communities often serve to reinforce moral values.

Topic 5: *divchyna* 'girl', *molodyi* 'young', *misiats* 'moon', *kupyty* 'to buy', *nosyty* 'to wear', *propyty* 'to drink away', *skhodyty* 'to go down', *ziity* 'to go up', *pustyty* 'to let go', *panna* 'lady', *puskaty* 'to let', *maty* 'mother', *vyity* 'to leave', *pyty* 'to drink', *vyno* 'wine', *druhyi* 'second', *kosa* 'braid', *zoria* 'star', *zelenyi* 'green', *vinok* 'wreath'.

This group centres on romantic and courtship themes. Actions like 'to buy' and 'to wear' suggesting the exchange of gifts. A 'girl' ('lady') asks her 'mother' to allow her to go on a date ('to let go', 'to leave') with a 'young' man. An unmarried girl often took pride in her 'braid', a cherished symbol of beauty. The 'wreath', intricately woven by the girl, held significant meaning – crafted either for her betrothed or for use in calendar-ritual festivities. During these rituals, the wreath could serve as an omen of her marital fate. The wreath also symbolizes the girl's purity and virginity. The recurring imagery of greenery ('green') marks the seasonal backdrop of these events, while mentions of celestial bodies suggest that courtship meetings took place from the evening ('to leave', 'moon') until sunrise ('to go up', 'star') – the only free time available after the day's labour. This thematic group frequently appears in wedding songs, family songs, and occasionally in calendar-ritual songs such as vesnianky or rusalka (songs associated with springtime and mermaid folklore), where the wreath served as a ritual object.

Topic 6: *vechir* 'evening', *zemlia* 'earth', *richka* 'river', *kin* 'horse', *kuvaty* 'to forge', *zabuty* 'to forget', *zozulia* 'cuckoo', *vyity* 'to leave', *skhylia* 'to bend', *nich* 'night', *poikhaty* 'to go away', *shchedryi* 'generous', *letity* 'to fly', *pokhylyty* 'to incline', *zhaty* 'to reap', *hora* 'hill', *mezha* 'boundary', *sidaty* 'to sit down', *vikontse* 'window', *maty* 'mother'.

The keywords associated with carols and shchedrivkas ('evening', 'generous', 'night', 'cuckoo', 'to fly') highlight their thematic connection to nature. This theme is further emphasized by landscape imagery such as earth', river, and 'hill'. The motif of Ukrainian agrarian life is reflected in terms like 'horse', 'to reap', and 'boundary'. Together, these elements underscore the profound relationship between humans and the natural world, capturing the rural landscapes in Ukrainian folklore.

Topic 7: *huliaty* 'to revel', *mylyi* 'beloved', *chornyi* 'black', *pyty* 'to drink', *sisty* 'to sit down', *brova* 'eyebrow', *dochka* 'daughter', *brat* 'brother', *vykhodyty* 'to come out', *tsvit* 'blossom', *chobit* 'boot', *kin*

'horse', *ity* 'to go', *bidnyi* 'poor', *vyidy* 'come out', *cherevychka* 'shoe', *shapochka* 'cap', *ruka* 'hand', *smert* 'death', *sestra* 'sister'.

The keywords in this topic are associated with ballads where a mother sells her daughter in a tavern ('to revel', 'to drink', 'daughter'). The daughter tries to escape an unwanted husband, running across impassable roads and injuring her feet ('blossom', 'boot'). However, he catches up to her on a 'horse', and in her despair, the 'poor' girl takes her own life ('death'). This group intricately weaves themes of familial relationships, including 'daughter', 'brother', and 'sister'. The songs in this group often serve as reflections on family life, highlighting its complexities and tensions. They emphasize the collective experience of grief and sorrow, a recurring motif in many traditional folk narratives.

Topic 8: *sad* 'garden', *staty* 'to stand', *dumaty* 'to think', *spaty* 'to sleep', *hroshi* 'money', *ity* 'to go', *molodyi* 'young', *nevistka* 'daughter-in-law', *vyshnevyi* 'cherry', *chas* 'time', *liahaty* 'to lie down', *pryity* 'to come', *temnyi* 'dark', *ruka* 'hand', *pora* 'time', *more* 'sea', *luhy* 'meadows', *zelenyi* 'green', *liubyty* 'to love', *khotity* 'to want'.

This topic is rich in words related to natural places, such as 'cherry' 'garden', 'sea', and 'green' 'meadows'. Another set of words focuses on actions or desires to act, including 'to stand', 'to think', 'to sleep', 'to go', 'to lie down', 'to come', 'to love', and 'to want'. Among the characters mentioned is the 'daughter-in-law'. This thematic group resonates with folk songs about familial relationships and songs of love, combining everyday activities with emotions and feelings.

Topic 9: *pity* 'to leave', *rano* 'early', *voda* 'water', *yisty* 'to eat', *kin* 'horse', *pshenytsia* 'wheat', *kudy* 'where', *kvitka* 'flower', *khlib* 'bread', *varyty* 'to cook', *lezhaty* 'to lie', *buty* 'to be', *mylyi* 'beloved', *vstaty* 'to rise', *nedilia* 'Sunday', *maty* 'mother', *sontse* 'sun', *natalko* 'Natalka', *hospodynia* 'hostess', *plyvty* 'to float'.

This theme encapsulates the rhythms of daily family life, featuring such persons as 'beloved', 'mother', 'Natalka', and 'hostess', with a strong focus on agricultural and domestic labour. Words like 'wheat' and 'bread' highlight the significance of staple foods in sustaining both life and tradition. Actions like 'to cook' and 'to eat' reflect familial activities, reinforcing the essential role of work and nourishment. This group emphasizes the interconnectedness of agricultural work and family life.

Topic 10: *liubyty* 'to love', *staty* 'to stand', *rosty* 'to grow', *vziaty* 'to take', *kalyna* 'viburnum', *divchyna* 'girl', *buty* 'to be', *smiiatysia* 'to laugh', *virno* 'faithfully', *chervonyi* 'red', *pity* 'to leave', *vysoko* 'high', *poikhaty* 'to go away', *derevo* 'tree', *kokhaty* 'to love', *boh* 'God', *ziat* 'son-in-law', *nahaika* 'whip', *khata* 'house', *ozhenytysia* 'to marry'.

This group combines themes of love, faith, and violence. One key aspect is the theme of songs about love ('faithfully', 'to love') and marriage ('to marry'). However, some songs explore the darker corners of family life. The 'whip' emerges as a symbol of cruelty, used by the 'son-in-law' to beat his wife, often to the point of death.

Topic 11: *dytyna* 'child', *staryi* 'old', *dolia* 'fate', *oko* 'eye', *malyi* 'small', *hai* 'grove', *kazaty* 'to say', *para* 'pair', *shchastia* 'happiness', *druhyi* 'second', *chornyi* 'black', *zhal* 'regret', *maty* 'mother', *lavka* 'bench', *divchyna* 'girl', *sluzhyty* 'to serve', *kara* 'punishment', *nashcho* 'why', *stan* 'figure', *dyvytysia* 'to look'.

This theme intertwines youth and age ('small', 'child', 'girl', 'mother', 'old'), emphasizing intergenerational relationships and the transmission of wisdom. Keywords like 'fate' and 'happiness' highlight existential concerns. One of the motifs involves a husband testing his wife's fidelity while he is away serving in the military. The wife, believing a letter falsely stating that her husband was leaving her, took her own life, leaving their 'small' 'child' in the care of 'old' grandfather. Upon learning the consequences of his test, the husband was consumed by 'black' 'regret'. This group reflects family life songs infused with moralistic and didactic messages.

Topic 12: *khlopets* 'boy', *hore* 'grief', *lykhyi* 'bad', *buvaty* 'to happen', *svit* 'world', *zdorov* 'hello', *myty* 'to wash', *divchyna* 'girl', *molodets* 'fine fellow', *dobre* 'well', *hodyna* 'hour', *harnyi* 'beautiful', *khotity* 'to want', *dodomu* 'homeward', *pryity* 'to come', *pane* 'lord', *zhyty* 'to live', *narobyty* 'to cause', *molod* 'young', *stoiaty* 'to stand'.

This group of keywords is dedicated to love songs, particularly focusing on the betrayal of a 'girl' by a 'fine fellow', 'boy'. He flirts with her, but quickly loses interest in her and 'to cause' her 'grief'. In

response, she curses him and refers to the moment of their meeting as a 'bad' 'hour'. Another motif in this group connected with the keyword 'lord', whose mention in songs typically carries a negative connotation. He is portrayed as someone who abuses his position and mistreats the peasants under his authority.

Topic 13: *khodyty* 'to walk', *stoiaty* 'to stand', *sestra* 'sister', *molodyi* 'young', *plakaty* 'to cry', *divchyna* 'girl', *voda* 'water', *dytyna* 'child', *khata* 'house', *podyvytysia* 'to look', *mylyi* 'beloved', *syn* 'son', *braty* 'to take', *buty* 'to be', *brat* 'brother', *zhyty* 'to live', *sertse* 'heart', *plach* 'weeping', *bilyi* 'white', *vdova* 'widow'

The keywords in this group focus on familial bonds ('sister', 'child', 'son', 'brother', 'widow') and romantic relationships ('young', 'girl', 'beloved', 'heart'). Centered on family connections, this theme highlights sibling relationships through terms like 'sister' and 'brother'. The frequent appearance of 'to cry' and 'weeping' adds an emotional layer, underscoring themes of sadness and loss. These motifs are central to love songs and songs about family life, reflecting both its joys and its sorrows.

Topic 14: *khata* 'house', *hist* 'guest', *kum* 'godfather', *kalyna* 'viburnum', *prykhodyty* 'to come', *porih* 'threshold', *luh* 'meadow', *boh* 'God', *robyty* 'to do', *noha* 'leg', *sim* 'seven', *znaty* 'to know', *mylyi* 'beloved', *pyty* 'to drink', *mynuty* 'to pass', *vorota* 'gate', *doroha* 'road', *lyst* 'letter', *vyity* 'to go out', *druhyi* 'second'.

Wedding rituals and the exchange of blessings ('God'), along with hospitality and social relationships, are at the core of this theme. 'Gate' and 'threshold' symbolize the wedding rituals through which the 'beloved' must overcome several obstacles to reach the bride. In wedding songs, the bride also expresses her gratitude, first to each member of her family, and then to the objects of her home. 'House' and 'guest' signify the importance of the home as a communal space, while 'godfather' highlights the familial bonds.

Topic 15: *maty* 'mother', *horilka* 'vodka', *myla* 'beloved', *dochka* 'daughter', *ity* 'to go', *boh* 'God', *khotity* 'to want', *baba* 'granny', *tudy* 'there', *batko* 'father', *syn* 'son', *pyty* 'to drink', *kozak* 'Cossack', *vrazhyi* 'enemy', *siudy* 'here', *molodyi* 'young', *pytatysia* 'to ask', *znaty* 'to know', *med* 'honey', *divchyna* 'girl'.

The keywords belong to songs about family life. This group emphasizes traditional family roles and generational relationships ('granny', 'father', 'son'). 'Mother' and 'daughter' with 'son' symbolize maternal bonds. Another motif belongs to love songs and mentions the 'enemy' 'Cossack', who prefers 'to drink' 'vodka' rather than go to his 'beloved' 'girl'. The pairing of the words 'vodka' and 'honey' is commonly used together to symbolize festivity and revelry.

Topic 16: *nemaie* 'absent', *divchyna* 'girl', *bilyi* 'white', *kozak* 'Cossack', *hora* 'hill', *did* 'grandpa', *sloza* 'tear', *pravda* 'truth', *chumak* 'Chumak', *postil* 'bedclothes', *zhdaty* 'to wait', *kin* 'horse', *vstavaty* 'to stand up', *durnyi* 'foolish', *mylyi* 'beloved', *more* 'sea', *kazaty* 'to say', *dolyna* 'valley', *znaty* 'to know', *buty* 'to be'

The keywords combine songs from various genres: chumak songs ('chumak'), love songs ('girl', 'Cossack', 'beloved'), and possibly some ballads. Chumaks, salt traders of old, embark on arduous journeys across the 'sea' and 'valley'. These themes capture longing and emotional absence. Words like 'absent' and 'tear' evoke feelings of emptiness and loss. The act of waiting ('to wait') underscores themes of hope amidst separation.

Topic 17: *daty* 'to give', *den* 'day', *ruchka* 'hand', *vil* 'ox', *napytysia* 'to get drunk', *vernysia* 'return', *korchma* 'tavern', *syn* 'son', *stil* 'table', *treba* 'need', *robyty* 'to do', *subota* 'Saturday', *pryity* 'to come', *bidnyi* 'poor', *braty* 'to take', *dodomu* 'homeward', *vmity* 'to know how', *lychko* 'face', *sestra* 'sister', *shyty* 'to sew'.

This group represents the songs of family life. The wealthy brother rejects his 'poor' 'sister'. He first invites her to visit him on 'Saturday', promising to share ('to give') his wealth with her, but instead, he acts in the opposite way. The keyword 'to sew' hints at the way the poor sister earns a living. Another theme in this group is revelry and the squandering of wealth ('to get drunk', 'to give', 'ox') in the 'tavern'.

Topic 18: *maty* 'mother', *ity* 'to go', *pity* 'to leave', *syn* 'son', *batko* 'father', *ridnyi* 'dear', *dribnyi* 'small', *doshch* 'rain', *dusha* 'soul', *nich* 'night', *divchyna* 'girl', *mylyi* 'beloved', *kladka* 'footbridge', *chornobryvyi*

'black-browed', *pryity* 'to come', *zasnuty* 'to fall asleep', *khmara* 'cloud', *selo* 'village', *porodyty* 'to give birth', *hudity* 'to vibrate'.

This group links weather patterns to emotional and familial themes ('mother', 'son', 'father'). Keywords like 'rain' and 'cloud' suggest metaphorical challenges or hardships. Terms such as 'dear', 'girl', 'beloved', and 'black-browed' are central to songs about love. An intriguing aspect of this theme is the juxtaposition of two oppositional words: 'to give birth' and 'to vibrate'. The latter, in the context of folk songs, refers to a husband beating his wife to death.

Topic 19: *mylyi* 'beloved', *pole* 'field', *divchyna* 'girl', *hraty* 'to play', *krasnyi* 'beautiful', *liuba* 'dear', *zdorovyi* 'healthy', *braty* 'to take', *holos* 'voice', *davaty* 'to give', *sydity* 'to sit', *kin* 'horse', *myla* 'beloved', *nazad* 'back', *khodyty* 'to walk', *lon* 'flax', *bachyty* 'to see', *mohty* 'to be able to', *diadko* 'uncle', *maiaty* 'to wave'.

This theme focuses on romantic encounters and playful courtship. Keywords such as 'beloved', 'girl', and 'dear' convey tenderness and affection. This group aligns with playful love songs, where natural landscapes often provide the backdrop for romantic meetings.

Topic 20: *pan* 'lord', *chuzhyi* 'stranger', *dvir* 'yard', *pole* 'field', *vziaty* 'to take', *dub* 'oak', *zhyto* 'rye', *chuty* 'to hear', *druzhyna* 'wife', *storona* 'side', *tsiluvaty* 'to kiss', *shukaty* 'to search', *tuman* 'fog', *vorona* 'crow', *zhaty* 'to reap', *kazaty* 'to say', *dyvo* 'miracle', *zhyty* 'to live', *liuli* 'lullaby', *dochka* 'daughter'.

This theme reflects the genres of a lullaby, as indicated by the keyword 'lullaby'. The term 'lord' suggests authority and affluence, highlighting social themes within the songs. Keywords such as 'wife' and 'daughter' relate to family life songs, while 'rye' and 'to reap' point to agrarian themes, particularly harvest songs.

After applying LDA to distinguish top 20 topics, each characterized by 20 keywords, a subset of the most significant terms was selected for further analysis. To enhance interpretability and focus on the core themes in Podillia region folk songs, the top 5 keywords for each topic, ranked by their beta values, were visualized in figure 2.

In figure 2 the x-axis represents the 20 topics, while the y-axis lists the top 5 keywords associated with each topic. Each cell's colour intensity reflects the beta value, which quantifies the strength of association between a keyword and a topic. A deeper blue shade indicates a stronger association, with beta values of 0.2 or higher showing the deepest blue. Gradations from light blue to white represent weaker associations, with cells corresponding to beta values approaching 0 remaining uncoloured.

A heatmap gives a sense of the relationship between keywords and clusters (e.g., which words are most associated with which clusters). The rows (keywords) and columns (topics) are clustered hierarchically, grouping similar words and topics together. This clustering of words and topics highlights relationships within and across topics, revealing thematic overlaps and distinctive patterns. It also demonstrates the shared and unique motifs that permeate the corpus, providing a deeper understanding of the thematic structure of Podillia region folk songs.

The next step of this research was to evaluate topic coherence. Without coherence evaluation, topics may consist of unrelated or vague keywords, making the results difficult to interpret or irrelevant to the research goals. Coherence ensures that topics are thematically robust and useful for analysis.

Figure 3 illustrates the coherence evaluation for the top 20 topics generated using LDA and applied to the Podillia region folk songs corpus. Four different metrics (Griffiths2004, CaoJuan2009, Arun2010, and Deveaud2014) are used to assess the quality of the topic modelling results. Each metric offers a unique perspective on the interpretability and quality of the identified topics:

- Griffiths2004 evaluates model likelihood, suggesting the number of topics that best balances model complexity and fit to the data. Higher likelihood scores indicate a better fit to the data.
- CaoJuan2009 measures topic redundancy by evaluating overlap, with lower values indicating less overlap among topics and better coherence, thus improving interpretability.
- Arun2010 examines the divergence between document-topic and topic-word distributions, with lower scores indicating better-defined and less ambiguous topics.
- Deveaud2014 focuses on semantic distinctiveness, where higher scores reflect greater separation and clarity among topics.

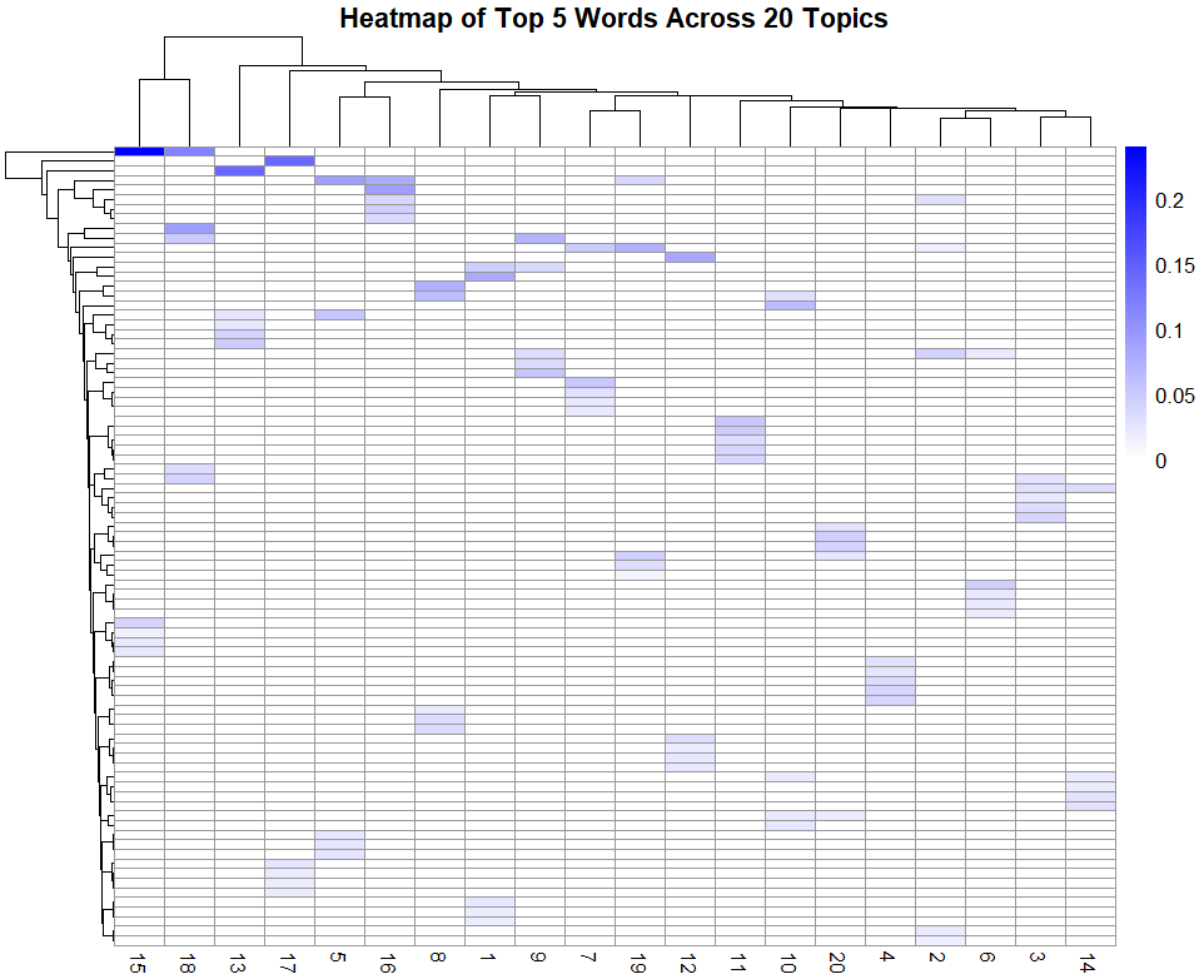

Figure 2: Heatmap of top 5 keywords across 20 topics in Podillia region folk songs.

Each metric provides a different perspective on the optimal number of topics:

1. Griffiths2004: The likelihood improves with an increasing number of topics but plateaus around 16-17 topics, indicating diminishing returns in model fit beyond this range.
2. CaoJuan2009: The values fluctuate between 20 and 7 topics but generally remain relatively low, suggesting acceptable levels of redundancy. Redundancy begins to increase significantly when the number of topics drops below 7, indicating merging or overlapping of themes.
3. Arun2010: The lowest scores occur after topics 15 and 16, implying these topics provide the best balance between distinctiveness and alignment between document-topic and topic-word distributions.
4. Deveaud2014: The metric suggests the highest semantic clarity and distinction are achieved at topic 7 (2.460026). However, coherence values remain relatively high before topic 7 and after it till topic 9-15, indicating this range also offers strong coherence.

While Deveaud2014 shows a preference for 7 topics, combining all metrics suggests that the range of 15-17 topics strikes the best balance between coherence, thematic granularity, and interpretability for the Podillia folk songs corpus. Topic 7 could be considered in cases where simpler, highly distinct thematic structures are preferred.

The results confirm that LDA quite effectively uncovers latent themes, with optimal configurations providing interpretable and culturally meaningful topics that align well with the underlying folkloristic content.

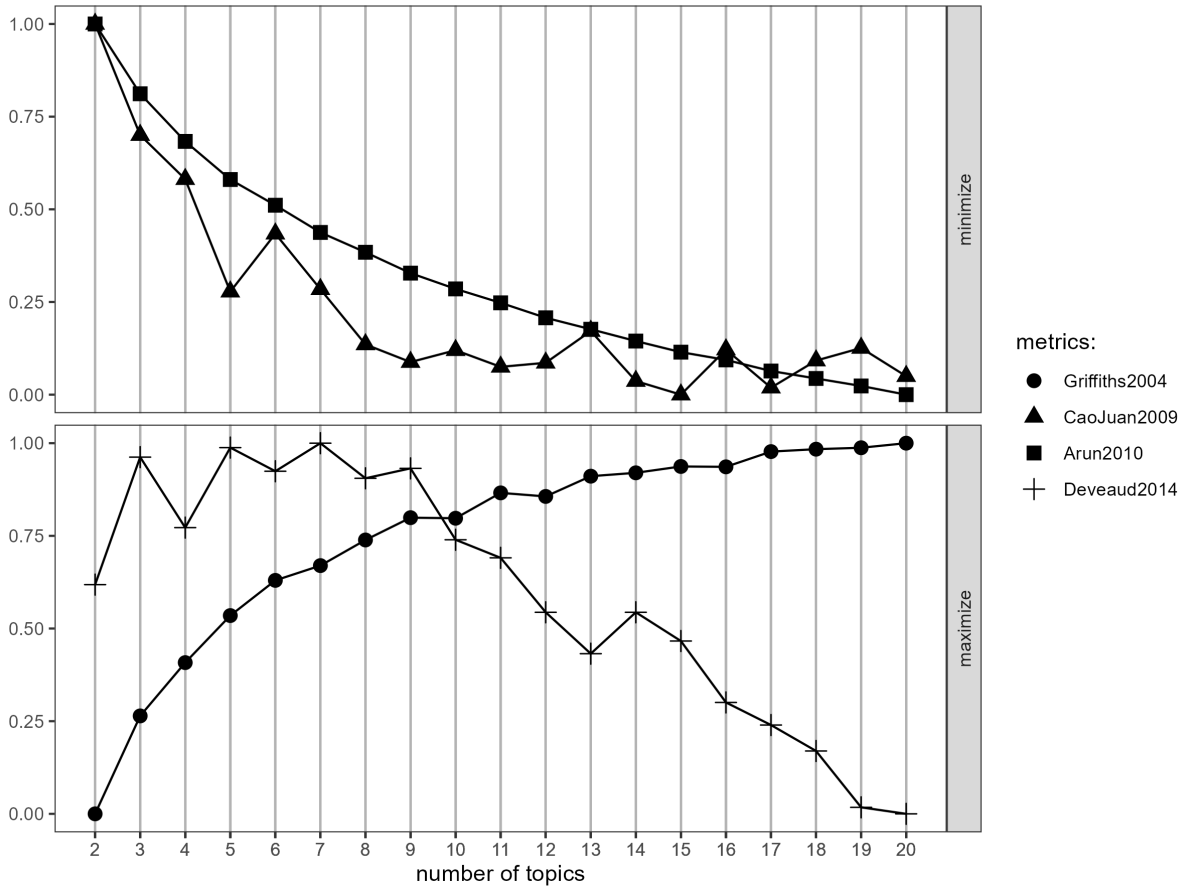

Figure 3: Topic coherence evaluation across metrics for Podillia region folk songs.

Building on the insights from LDA, further analysis using word embeddings and clustering offers a deeper exploration of semantic relationships within the corpus. This analysis of Podillia region folk songs using word embeddings and clustering highlights significant semantic patterns, reflecting cultural motifs and themes. PCA statistics and clustering provide a quantifiable basis for interpreting folk motifs.

Principal components are linear combinations of the original features (in this case, word embedding dimensions) that maximize the variance captured in each successive dimension. They serve as a dimensionality reduction tool, summarizing the complex relationships between words in a simpler, interpretable form.

Keywords are grouped by their cluster membership, as determined by K-means clustering of word embeddings. Each cluster represents a thematic group of semantically related words. Figure 4 shows word clusters generated from word embeddings, focusing on the most representative 20 keywords per cluster. The principal components (PC1 and PC2) derived from PCA provide a two-dimensional projection of the high-dimensional word embedding space. Bubble sizes correspond to term frequencies, emphasizing the relative importance of keywords within each cluster.

The 20 clusters reveal distinct thematic groups. While the keywords identified here differ slightly from those distinguished by LDA, they largely align and share similar key motifs, including familial roles, emotions, nature, agricultural work, and traditions.

The range of PC1 and PC2 values (-4 to 4) represents the variation in the dataset. Larger ranges for PC1 and PC2 indicate greater diversity in the dataset's semantic dimensions. Words closer to the origin (0,0) are less distinct in their thematic contribution, while those farther away contribute strongly to the thematic dimensions.

Cluster analysis highlights a focus on family, emotional depth, and an emphasis on nature and calendar rituals. Central themes include familial relationships ('mother', 'daughter', 'son'), reflecting the

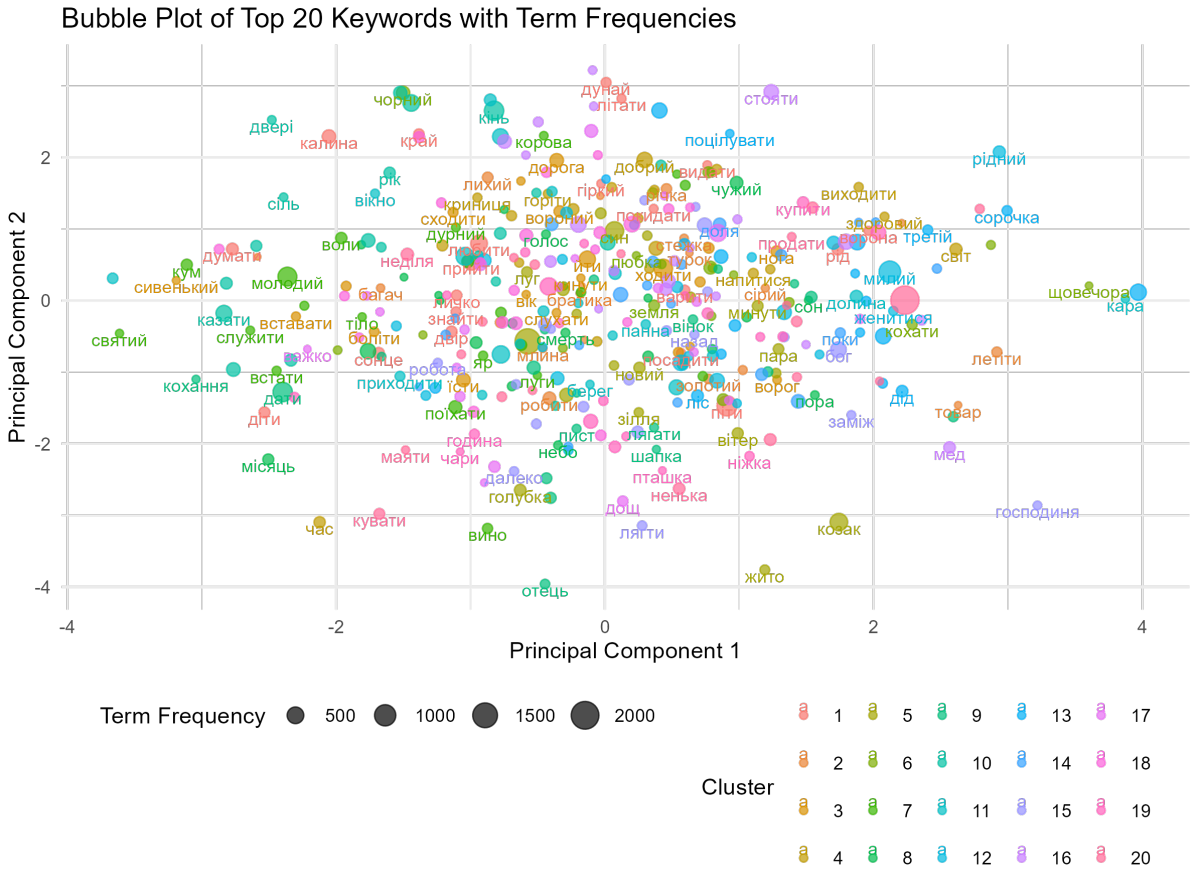

Figure 4: Visualization of top 20 keywords per cluster in Ukrainian folk songs using word embeddings and PCA. Keywords are grouped into clusters representing thematic motifs, with bubble size indicating term frequency within the corpus. The x-axis (PC1) and y-axis (PC2) represent the two principal components, capturing the largest variances in the semantic relationships among keywords.

importance of kinship and community in Ukrainian culture. Clusters with terms such as 'love', 'grief', and 'heart' capture the emotional richness of Ukrainian folk song narratives.

5. Conclusion

This study has demonstrated the potential of topic modelling in uncovering the thematic richness of Podillia region folk songs. By applying LDA to a Ukrainian-language corpus, the research provides a scalable and systematic approach to analysing the cultural and narrative elements embedded in these songs. The results reveal top 20 distinct thematic clusters, which offer a nuanced view of recurring motifs and narrative structures present within the folk songs. These include themes of family life, romantic relationships, agrarian labour, ritual practices, and emotional struggles.

Topics such as 1, 4, and 9 emphasize the challenges and dynamics of family relationships, including maternal care, sibling bonds, and marital conflicts. Themes like 2, 5, and 12 capture the spectrum of romantic relationships, from playful courtship and deep affection to betrayal and heartbreak. Wedding rituals are a key focus, with several topics capturing different aspects of these ceremonies. Keywords such as 'gate' and 'threshold' reflect the physical and symbolic transitions during weddings, while familial blessings and exchanges of gratitude emphasize the communal nature of these events (topic 14). Other ritualistic elements appear in carols and shchedrivkas, where themes of nature and abundance dominate (topic 6). The agrarian and rural themes prevalent in 1, 6, and 20 reveal the centrality of labour and nature in traditional life.

The findings from LDA align well with traditional folkloristic classifications while also offer new insights and refine existing understandings of Podillia folk traditions. Many of the thematic clusters identified through LDA, such as family bonds, chumak's journey, love, and the harvest are consistent with previous folkloristic research. However, by leveraging computational methods, this study uncovers more granular and nuanced semantic patterns that might have been overlooked in previous qualitative analyses. For instance, it reveals contrasting motifs within a single theme, such as giving birth to a child versus a husband beating his wife to death. The combination of LDA with PCA and clustering techniques deepens the understanding of these traditional narratives, highlighting the interconnectedness of various cultural motifs within Podillia folk songs.

The application of computational methods to the Podillia region folk song corpus presented several challenges, particularly due to the linguistic complexity and cultural context of the material. One of the primary obstacles was dealing with the linguistic complexity of the Ukrainian language, including its morphology and regional dialects. To address this, custom preprocessing steps were developed, such as tokenization and lemmatization using the Ukrainian UDPipe model, to better handle the corpus's linguistic features. Another challenge was ensuring that the semantic richness of the folk songs was captured accurately. The use of word embeddings and PCA helped mitigate this issue by mapping words to dense vector spaces, facilitating the identification of meaningful relationships between terms. Moreover, aligning computational findings with traditional folkloristic interpretations required a careful balance of qualitative insights and quantitative methods. Future research could address these challenges by incorporating multilingual models such as BERT or BERTopic to enhance the contextual analysis and explore comparative studies between Podillia and neighbouring regions and/or try to apply translation into English for comparing with folk songs of another country, offering deeper insights into shared or divergent cultural motifs.

The results confirm that LDA effectively uncovers the latent themes in the Podillia folk songs corpus, with the optimal configuration (between 15-17 topics) providing both interpretability and thematic coherence. The topic coherence evaluation, using four different metrics (Griffiths2004, CaoJuan2009, Arun2010, and Deveaud2014), highlights the robustness of the LDA model in providing meaningful topics that align well with the folkloristic content.

This study highlights the possibilities of combining computational methods with cultural analysis to study oral traditions, offering a scalable approach to understanding the complex narratives within Ukrainian folk songs. Building on LDA, further analysis using word embeddings and clustering techniques adds a deeper layer of understanding. PCA and clustering revealed semantic patterns that reflect the central role of family, emotion, nature, and traditions in Podillia folk songs. These computational tools bridge the gap between qualitative folklore studies and quantitative text mining, offering a holistic view of the thematic and narrative structure of Ukrainian folk songs.

Thus, this research demonstrates that topic modelling through LDA, combined with word embeddings and clustering, and connected with traditional folkloristic classifications, paves the way for a deeper understanding of Ukrainian cultural heritage, while also advancing the methodology of computational folkloristics.

Data Availability Statement: The text corpus and R scripts to analyse the topic modelling of Podillia region folk songs are available at Zenodo [22].

Funding: This work was supported by the Estonian Research Council (project MOB3JD1218 Traditional Estonian and Ukrainian Folksongs: Comparative Corpus-Based Computational Analysis), and the Ministry of Education and Research (research projects TK215 Estonian Roots: Centre of Excellence for transdisciplinary studies on ethnogenesis and cultural diversity).

Declaration on Generative AI: During the preparation of this work, the author used Grammarly for Chrome (Grammarly's browser extensions) to improve the article's grammar. The author also employed ChatGPT-4o to proofread several passages and suggest improved formulations. However, not all suggestions were accepted, and the author made the final decisions regarding the text. The author reviewed and edited all content as needed and takes full responsibility for the publication's content.

References

[1] K. Kuutma, Cultural identity, nationalism and changes in singing traditions, Folklore. Electronic Journal of Folklore 2 (1996) 124–141. doi:10.7592/fejf1996.02.ident.

[2] M. Stokes (Ed.), Ethnicity, Identity and Music: The Musical Construction of Place, BERG Ethnic Identities Series, Oxford/Providence, USA, 1994.

[3] E. M. Thompson, Nationalism, Imperialism, Identity: Second Thoughts, Modern Age 40 (1998) 250–261. URL: https://modernagejournal.com/nationalism-imperialism-second-thoughts/229904/.

[4] J. Collins, P. Long, 'Fillin' in Any Blanks I Can': Online Archival Practice and Virtual Sites of Musical Memory, in: Sites of Popular Music Heritage: Memories, Histories, Places, Routledge, New York and London, 2014.

[5] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (2012) 77–84. doi:10.1145/2133806.2133826.

[6] A. Murakami, P. Thompson, S. Hunston, D. Vajn, 'What is this corpus about?': using topic modelling to explore a specialised corpus, Corpora 12 (2017) 243–277. doi:10.3366/cor.2017.0118.

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

[8] O. Petrovych, I. Zavalniuk, V. Bohatko, Exploring the Semantics and Structure of Vocatives in Ukrainian Folk Songs, Folklore: Electronic Journal of Folklore 94 (2024) 233–266. doi:10.7592/fejf2024.94.ukrainian.

[9] R. Egger, Topic Modelling: Modelling Hidden Semantic Structures in Textual Data, in: R. Egger (Ed.), Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications, Springer International Publishing, Cham, 2022, pp. 375–403. doi:10.1007/978-3-030-88389-8_18.

[10] O. Sobchuk, A. Šeļa, Computational thematics: comparing algorithms for clustering the genres of literary fiction, Palgrave Communications 11 (2021) 438. doi:10.1057/s41599-024-02933-6.

[11] E. Nylander, D. Holmer, The latent structure of educational offerings—tracing topics from folk high school catalogues through large-scale content analyses, Zeitschrift Für Weiterbildungsforschung 45 (2022) 295–319. doi:10.1007/s40955-022-00222-w.

[12] M. D. Devi, N. Saharia, Exploiting Topic Modelling to Classify Sentiment from Lyrics, in: A. Bhattacharjee, S. K. Borgohain, B. Soni, G. Verma, X.-Z. Gao (Eds.), Machine Learning, Image Processing, Network Security and Data Sciences, volume 1241 of *Communications in Computer and Information Science*, Springer Singapore, Singapore, 2020, pp. 411–423. doi:10.1007/978-981-15-6318-8_34.

[13] K. Dakshina, R. Sridhar, LDA Based Emotion Recognition from Lyrics, in: M. Kumar Kundu, D. P. Mohapatra, A. Konar, A. Chakraborty (Eds.), Advanced Computing, Networking and Informatics-Volume 1, volume 27 of *Smart Innovation, Systems and Technologies*, Springer International Publishing, Cham, 2014, pp. 187–194. doi:10.1007/978-3-319-07353-8_22.

[14] E. Laoh, I. Surjandari, L. R. Febirautami, Indonesians' Song Lyrics Topic Modelling Using Latent Dirichlet Allocation, in: 2018 5th International Conference on Information Science and Control Engineering (ICISCE), 2018, pp. 270–274. doi:10.1109/ICISCE.2018.00064.

[15] P. Wanjantuk, N. Pinitkarn, S. Sengthavideth, J. Matthayomnan, A. Meesomboon, Unveiling Emotions and Themes in Thai Songs via Topic Modeling, in: 2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE), 2024, pp. 682–688. doi:10.1109/JCSSE61278.2024.10613629.

[16] S. Panda, V. P. Namboodiri, S. T. Roy, Visualizing Music Genres using a Topic Model, 2021. URL: https://arxiv.org/abs/2103.00127. arXiv:2103.00127.

[17] M. Sarv, Regilaulude teema-analüüs: võimalusi ja väljakutseid [Topic analysis of Estonian runosongs: Prospects and challenges], Methis Studia Humaniora Estonica 21 (2020). doi:10.7592/methis.v21i26.16914.

[18] G. Strle, M. Marolt, Language Technologies in Humanities: Computational Semantic Analysis in

Folkloristics, in: Conference on Language Technologies & Digital Humanities, Ljubljana, 2016, 2016, pp. 227–229. doi:10.5281/zenodo.14165156.

[19] P. Kryndach, V. Vysotska, S. Chyrun, L. Chyrun, S. Goloshchuk, R. Holoshchuk, Analysis of Semantic Relationships in Ukrainian Text Content Based on Word2Vec and Machine Learning, in: 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), 2023, pp. 1–6. doi:10.1109/CSIT61576.2023.10324074.

[20] N. Khairova, Y. Holyk, D. Sytnikov, Y. Mishcheriakov, N. Shanidze, Topic Modelling of Ukraine War-Related News Using Latent Dirichlet Allocation with Collapsed Gibbs Sampling, in: V. Vysotska, Y. Burov (Eds.), Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems. Volume III: Intelligent Systems Workshop, Lviv, Ukraine, April 12-13, 2024, volume 3688 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 1–15. URL: https://ceur-ws.org/Vol-3688/paper1.pdf.

[21] A. Verbytska, Topic modelling as a method for framing analysis of news coverage of the Russia-Ukraine war in 2022–2023, Language & Communication 99 (2024) 174–193. doi:10.1016/j.langcom.2024.10.004.

[22] O. Petrovych, Applying Topic Modelling to the Folk Song Corpus of the Podillia Region, 2024. doi:h10.5281/zenodo.14281803.

[23] O. Dei (Ed.), Pisni Yavdokhy Zuikhy: zapysav Hnat Tantsiura [Songs of Yavdokha Zuikha: recorded by Hnat Tantsiura], Naukova dumka, Kyiv, 1965.

[24] L. Yefremova, M. Dmytrenko (Eds.), Narodna pisni Khmelnychchyny (z kolektsii zbyrachiv folkloru) [Folk songs of Khmelnytskyi region (from the collections of folklore collectors)], Naukova dumka, Kyiv, 2014.

[25] S. Myshanych (Ed.), Pisni Podillia: zapysy Nasti Prysiazhniuk v seli Pohrebyshche. 1920-1970 rr. [Songs of Podillia: recordings of Nastia Prysiazhniuk in the village of Pohrebyshche. 1920-1970.], Naukova dumka, Kyiv, 1976.

[26] G. Diaz, A. Suriyawongkul, M. Pukhalskyi, B. Solomon, Stopwords ISO, 2020. URL: https://github.com/stopwords-iso/stopwords-iso.

[27] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of Machine Learning Research 3 (2003) 993–1022. doi:10.5555/944919.944937.

[28] M. Straka, J. Straková, Universal Dependencies 2.0 Models for UDPipe (2017-08-01), 2017. URL: http://hdl.handle.net/11234/1-2364.

[29] M. Steyvers, T. Griffiths, Probabilistic topic models, in: T. K. Landauer, D. S. McNamara, S. Dennis, W. Kintsch (Eds.), Handbook of latent semantic analysis, Lawrence Erlbaum Associates Publishers, 2007, pp. 427–448. doi:10.4324/9780203936399.ch21.

[30] M. Röder, A. Both, A. Hinneburg, Exploring the Space of Topic Coherence Measures, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM '15, Association for Computing Machinery, New York, NY, USA, 2015, p. 399–408. doi:10.1145/2684822.2685324.

[31] T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National Academy of Sciences 101 (2004) 5228–5235. doi:10.1073/pnas.0307752101.

[32] J. Cao, T. Xia, J. Li, Y. Zhang, S. Tang, A density-based method for adaptive LDA model selection, Neurocomputing 72 (2009) 1775–1781. doi:10.1016/j.neucom.2008.06.011, Advances in Machine Learning and Computational Intelligence.

[33] R. Arun, V. Suresh, C. E. Veni Madhavan, M. N. Narasimha Murthy, On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations, in: M. J. Zaki, J. X. Yu, B. Ravindran, V. Pudi (Eds.), Advances in Knowledge Discovery and Data Mining, volume 6118 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 391–402. doi:10.1007/978-3-642-13657-3_43.

[34] R. Deveaud, E. Sanjuan, P. Bellot, Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval, Document numérique - Revue des sciences et technologies de l'information. Série Document numérique (2014) 61–84. URL: https://hal.science/hal-01002716. doi:10.3166/DN.17.1.61-84.

[35] J. Pennington, R. Socher, C. Manning, GloVe: Global Vectors for Word Representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: https://aclanthology.org/D14-1162. doi:10.3115/v1/D14-1162.

[36] J. Chuang, C. D. Manning, J. Heer, Termite: visualization techniques for assessing textual topic models, in: Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI '12, Association for Computing Machinery, New York, NY, USA, 2012, p. 74–77. doi:10.1145/2254556.2254572.