Jurnal Teknik Informatika Vol. 18 No. 2, October 2025 (248-256)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.46643

JURNAL TEKNIK INFORMATIKA

Homepage : http://journal.uinjkt.ac.id/index.php/ti

Uncovering Hidden Themes in Indie Music: Crisp-Dm Guided LDA Topic Modeling on A Kaggle-Based Lyric Generation Dataset

Thoyyibah T [1]*, Yan Mitha Djaksana [2]

[1]Information System Management Department, BINUS Graduate Program-Master of Information System Management, University of Bina Nusantara
[2]Information Technology Program, Faculty of Computer Science, Pamulang University
[1]Jl. Kebon Jeruk Raya No. 27, Jakarta Barat, 11530 Indonesia
[2]Jl. Raya Puspitek, Buaran, Kec. Pamulang, Kota Tangerang Selatan, Banten, 15310 Indonesia

Article:
Accepted: August 30, 2025
Revised: April 22, 2025
Issued: October 30, 2025

***Correspondence Address:**
thoyyibah.t@binus.ac.id

ABSTRACT

The development of music has produced many works in the form of data, especially lyrical data, which provide insight into the semantic structure of music. This study explores latent thematic patterns in the indie lyric dataset from Kaggle by applying Latent Dirichlet Allocation (LDA), which is the first LDA study of indie music lyrics in the Indonesian context with the interpretation of love, emotional needs, romance, and inner conflict. The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology can be effectively applied to unstructured data, opening up opportunities for better music classification. The methodological stages include business and data understanding, data preparation, modelling, evaluation, and dissemination. In the early stages, the Kaggle dataset implemented Natural Language Processing, which was done with case folding, punctuation removal, stopword removal, stemming, and tokenization. The LDA model is trained by identifying five topics with different interpretations. Visualization in WordClouds, with topic distribution on datasets and title-based topic mapping. This model yielded a coherence value of 0.3044, which indicates limited semantic consistency, which means the words in the topic have a reasonably good relationship, but there is still potential for refinement in subsequent studies. The limitations of this study include the limited size of the dataset, with only 347 rows and slight variation in interpretation. For future research, it is recommended to use larger datasets and more diverse interpretations and apply more machine learning models.

Keywords : *natural language processing; topic modeling; CRISP-DM; LDA; lyrics dataset.*

1. INTRODUCTION

Music is a sound called art[1] [2][3]. Music has so many benefits. One of the benefits of music is to reduce stress and relate to emotions [4][5][6][7]. Music is closely related to the data of the lyrics or the accompanying text [8][9][10].

Indie music, including mainstream music, has poetic, experimental lyrics that characterize the deep feelings of its creator. Indie music provides a vast space with artistic diversity. Indie music offers a thematic richness that researchers have not explored with a computational approach. However, this semantic complexity is also a challenge in the automated analysis, modelling, and learning from raw data using Natural Language Processing (NLP) approaches) [11].

Exploring topics in the lyrics has its challenges that can be processed, such as complex song structures with elements of repetition. In addition, the song's structure, depending on the genre, adds a layer of complexity to the song [12][13]. Another challenge is the varying length of song lyrics and ambiguous meanings found in song lyrics [14]. Different genres of music will also produce different lyrical structures [15]. The quality of annotations with the song sampling method must be analyzed further [16].

Machine learning is increasingly developing, with one of the methods often used to identify hidden themes in song, namely Latent Dirichlet Allocation (LDA) [17]. The data set is divided into several topics, which results in a coherence value. [18]. LDA is very effective in finding different topics in big data. Topics contain many words with multinomial distribution [19] [20].

Probabilistic-based LDA includes topic modelling algorithms to determine several topics from a group of words without manual Supervision [18]. Songs with rich lyrics can use LDA to identify repeated hidden themes, such as freedom, love, criticism, and hope, which are spread throughout the song's lyrics.

Several studies have used LDA and NLP in music or lyrics, including Mood classification in lyrics with LDA and XGBoost model [21]. Sentiment using LDA with an analysis of 150 song lyrics [22]. Probability of the topic applying LDA from the lyric corpus [23]. LDA was also used to analyze the historical evolution of German popular music lyrics [24]. Jingju's corpus of lyrics in Peking opera was once also used with topic modelling [25]. Sentiment analysis with correlation models and LDA evaluation was conducted in various domains, including music [26].

Although many experiments on LDA have been carried out, for example, analysis of news reviews, product reviews, and social media comment reviews, applications in the domain of music, especially indie songs, have never been carried out, let alone LDA in the context of the Indonesian language.

Therefore, this study aims to apply LDA to indie music lyrics as an initial step toward deeper semantic archery in contemporary music. This study uses the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology to ensure the process is carried out systematically. CRISP-DM has business understanding, data understanding, data preparation, modelling, evaluation, and deployment [27]. This methodology is very flexible, with stages used to deal with various types of projects. [28] [29][30] [31].

This methodology is also very suitable for handling unstructured data such as text in song lyrics. This study uses a free dataset from Kaggle called the Lyrics Generation dataset, consisting of 347 song lyrics. This dataset allows for extensive analysis of the themes or topics contained in all song lyrics. The primary focus lies on how the themes contained in the lyrics can be visualized and their coherence calculated.

The application of LDA using the CRISP-DM methodology can contribute to understanding the thematic structure of song lyrics so that these findings can be used by researchers in the fields of music, artists, and computer science by implementing AI. Overall, this study shows how NLP can be used on song lyric text data and introduces the CRISP-DM framework to analyze data systematically. Utilizing LDA, this research can open new insights into the multidisciplinary field between computer science, linguistics, and music.

2. METHODS

This study uses the CRISP-DM methodology. CRISP-DM is a systematic methodology with several phases [32] [33]. This study uses the CRISP-DM methodology.

CRISP-DM is a systematic methodology with several phases. [34].

Figure 1 shows the five stages of CRISP-DM that are carried out. Only this study is carried out up to the evaluation stage. These stages include business understanding and data understanding, data preparation, modelling, and evaluation. [35] [36].

This stage consists of business understanding and data understanding, data preparation, modelling, and evaluation. The dataset used with the https://www.kaggle.com/datasets/pratiksaha198/lyrics-generation link is the result of lyrics using Recurrent-Neural-Networks (RNN) from the dataset created by scraping the GENIUS website using its API.

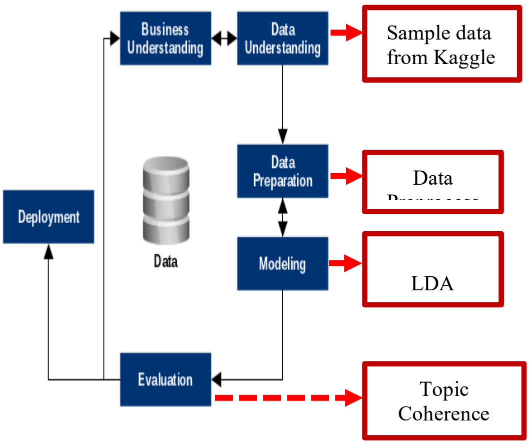


Figure 1. *Crisp-DM Method*

The libraries used for the initial steps are Pandas and numpy for data manipulation, numerical calculations, and cleaning and preparing text data. Other libraries are the Dictionary Corporation, which creates a dictionary from text, and LdaModel, which builds LDA topic models. The coherence Model is used to measure coherence, and WordCloud is used to create a word cloud that describes the frequency of words in a topic.

The second stage is text data preprocessing. At this stage, the process is carried out, and the initial processing of text data is carried out using case folding, removing punctuation, removing stopwords, stemming, and tokenization. Case folding is done by changing the text to lowercase. Removing punctuation is done by removing numbers, punctuation, etc., for example, 1-9!@#$%^&*(). Removing stop words is done by removing words that often appear but do not have any information, such as pronouns, prepositions, conjunctions, etc. Stemming is done by transforming words to produce basic words.

Gensim LdaModel is the model used in this research. The number of topics is five based on initial exploration with consideration of optimal topic interpretability. The model is trained through 100 iterations and 10 passes to ensure the convergence stability of the topic distribution. The alpha parameter is set to an auto value that allows the model to self-adjust according to the distribution of topics in the document based on the complexity of the data. A balanced topic in the document, while ETA (or beta) is set to 'auto', allows the model to adjust the distribution of words in the topic adaptively. In addition, per_word_topics=True needs to be enabled to generate a topic distribution at the word level,

allowing for a more in-depth analysis of the contribution of words in each topic. The metric used is the Coherence Score using c_v, which measures semantic similarity between words in one topic.

The third stage is LDA preprocessing. This stage is carried out by forming a corpus from the lyrics data. A corpus represents a document in the form of numbers based on how many words appear, while the dictionary stores the index of each word that is considered unique. The LDA model is formed from the converted data as its primary input. The LDA model is trained with the Gensim library. This library provides distributed learning of the topics in the document. The training process with iteration is expected to produce a stable model. After that, the model will produce some topic lists from the existing data set. Each song will produce a proportion of songs based on the dominant theme. This stage is the core stage because the analysis of topics with hidden semantic structures begins to appear.

The fourth stage is the last carried out in this study. At this evaluation stage, the quality of the model is measured using the coherence score metric. This coherence score is needed to analyze and measure the extent to which words in the topic are semantically related. This evaluation is carried out to measure whether the resulting topic is mathematically accurate and linguistically reasonable. The Coherence value is calculated by comparing the occurrence of words in relevant documents. The higher the

ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v18i2.46643

coherence value produced, the better the model quality used.

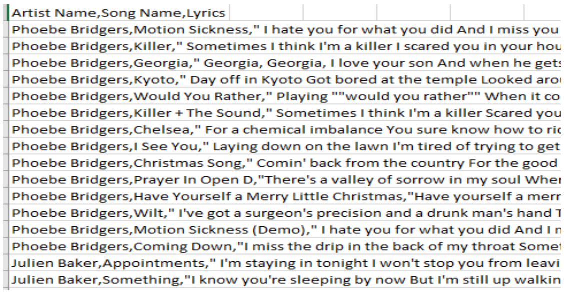
This study uses an alternative model, Non-negative Matrix Factorization (NMF), as a comparison. This model is part of matrix factorization that is deterministic and not probabilistic-based, in contrast to Latent Dirichlet Allocation (LDA), which uses generative and probabilistic approaches in topic formation. In NMF, a document-word matrix is decomposed into two non-negative matrices: one represents the document's relationship to the topic, and the other represents the topic's relationship to the word. The advantage of NMF lies in the interpretability of the results because the limitations of non-negative make the results easier to interpret intuitively, for example, as the "contribution" of a word to a topic.

3. RESULTS AND DISCUSSION

The results and discussion are a development of the steps in the research method.

3.1. Business Understanding and Data Understanding

Business Understanding and Data Understanding is done by preparing the dataset to be processed. The dataset in Figure 2 consists of 3 columns, namely artist name, song name, and lyrics, with 347 rows. The artist's name is the name of the singer who sang the song, such as Phoebe Bridgers. The song name is the title of the song sung by the singer, for example, Motion Sickness.

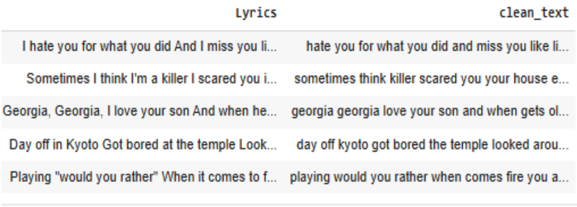


```
Artist Name,Song Name,Lyrics
Phoebe Bridgers,Motion Sickness," I hate you for what you did And I miss you
Phoebe Bridgers,Killer," Sometimes I think I'm a killer I scared you in your hou
Phoebe Bridgers,Georgia," Georgia, Georgia, I love your son And when he gets
Phoebe Bridgers,Kyoto," Day off in Kyoto Got bored at the temple Looked aro
Phoebe Bridgers,Would You Rather," Playing ""would you rather"" When it co
Phoebe Bridgers,Killer + The Sound," Sometimes I think I'm a killer Scared you
Phoebe Bridgers,Chelsea," For a chemical imbalance You sure know how to ri
Phoebe Bridgers,I See You," Laying down on the lawn I'm tired of trying to get
Phoebe Bridgers,Christmas Song," Comin' back from the country For the good
Phoebe Bridgers,Prayer In Open D,"There's a valley of sorrow in my soul Whe
Phoebe Bridgers,Have Yourself a Merry Little Christmas,"Have yourself a merr
Phoebe Bridgers,Wilt," I've got a surgeon's precision and a drunk man's hand T
Phoebe Bridgers,Motion Sickness (Demo)," I hate you for what you did And I r
Phoebe Bridgers,Coming Down,"I miss the drip in the back of my throat Some
Julien Baker,Appointments," I'm staying in tonight I won't stop you from leavi
Julien Baker,Something,"I know you're sleeping by now But I'm still up walkin
```

Figure 2. *Dataset* (https://www.kaggle.com/datasets/pratiksaha198/lyrics-generation.)

3.2. Data Preparation

Data Preparation in Figure 3 is done by processing text data initially. This stage is done by case folding, removing punctuation, removing stopwords, stemming, and tokenization. The results of the cleaned data are seen in Figure 3 with the clean column.



Lyrics	clean_text
I hate you for what you did And I miss you li...	hate you for what you did and miss you like li...
Sometimes I think I'm a killer I scared you i...	sometimes think killer scared you your house e...
Georgia, Georgia, I love your son And when he...	georgia georgia love your son and when gets ol...
Day off in Kyoto Got bored at the temple Look...	day off kyoto got bored the temple looked arou...
Playing "would you rather" When it comes to f...	playing would you rather when comes fire you a...

Figure 3. *Text dataset cleaning*

3.3. Modelling

Modelling is done by implementing the LDA model on the dataset used. Then, the visualization in the form of a word cloud of processed lyrics, as in Figure 4, and the Worldcloud per topic, as in Figure 5, will be displayed. The distribution of topics in song lyrics in the form of a pie chart is also done as in Figure 6 with the results of 18.5% for topic 1, 22.9% for topic 2, 16.9% for topic 3, 22.2% for topic 4, 19.5% for topic 5. Figure 7 shows the number of documents per topic, consisting of 67 documents for topic 1, 74 for topic 2, 60 for topic 3, 78 for topic 4, and 68 for topic 5. Figure 8 is the distribution of topics based on song titles.

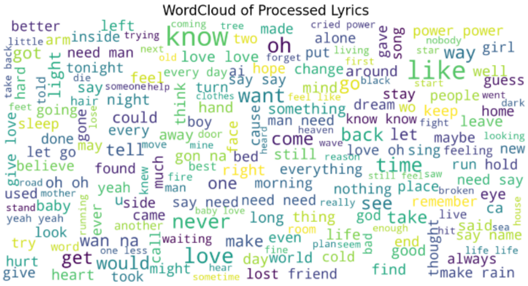


Figure 4. *WorlCloud lyrics on dataset*

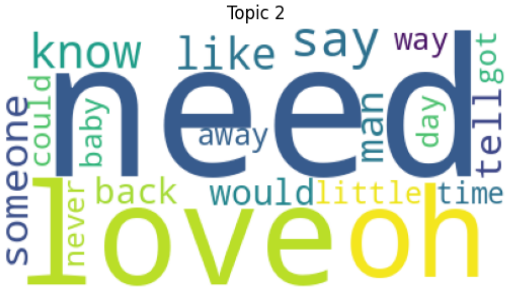


Figure 5. *Distribution of lyrics on topics*

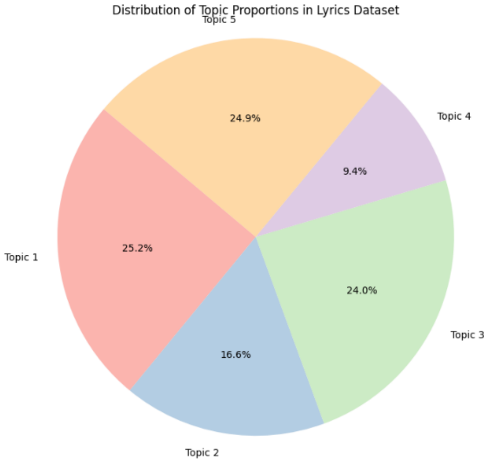


Figure 6. *Distribution of topic proportions*

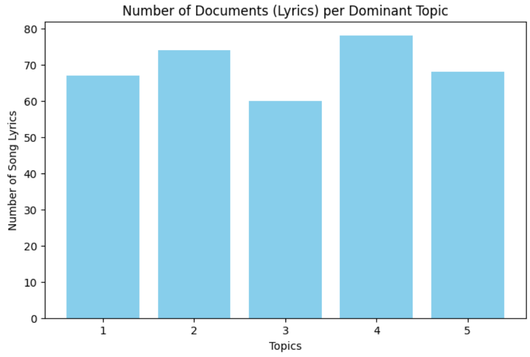


Figure 7. *Number of documents per topic*

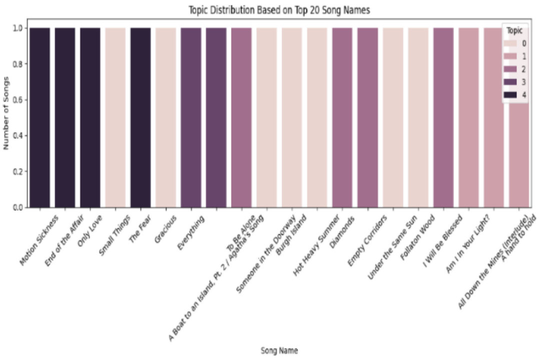
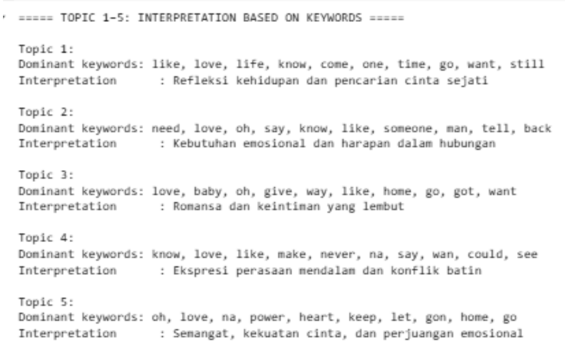


Figure 8. *Topic Distribution by 20 Song Title*

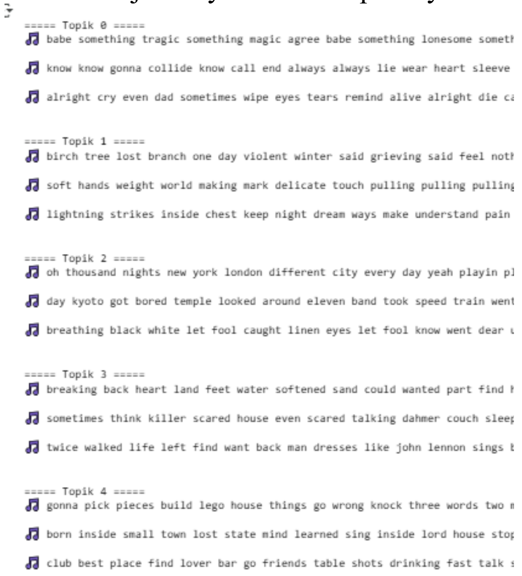
Figure 9 is an interpretation based on topic one keywords: love, life, know, come, one, time, go, want, and still, resulting in a reflection of life and the search for true love. Topic 2 dominant keywords are need, love, oh, say, know, like, someone, man, tell, and back with the interpretation of emotional needs and expectations in relationships. Topic 3 with dominant keywords, namely love, baby, oh, give, way, like, home, go, got, and want, produces an interpretation of romance and gentle intimacy. Topic 4 has dominant keywords: know, love, like, make, never, na, say, wan, could, and see, producing an interpretation of deep feelings and inner conflict. Topic 5 with dominant keywords, namely oh, love, na, power, heart, keep, let, going to, home, and go, produces an interpretation of enthusiasm, the power of love, and emotional struggle.

Figure 9. *Interpretation Based on Keywords*



```
===== TOPIC 1-5: INTERPRETATION BASED ON KEYWORDS =====

Topic 1:
Dominant keywords: like, love, life, know, come, one, time, go, want, still
Interpretation     : Refleksi kehidupan dan pencarian cinta sejati

Topic 2:
Dominant keywords: need, love, oh, say, know, like, someone, man, tell, back
Interpretation     : Kebutuhan emosional dan harapan dalam hubungan

Topic 3:
Dominant keywords: love, baby, oh, give, way, like, home, go, got, want
Interpretation     : Romansa dan keintiman yang lembut

Topic 4:
Dominant keywords: know, love, like, make, never, na, say, wan, could, see
Interpretation     : Ekspresi perasaan mendalam dan konflik batin

Topic 5:
Dominant keywords: oh, love, na, power, heart, keep, let, gon, home, go
Interpretation     : Semangat, kekuatan cinta, dan perjuangan emosional
```

In the results of the topic modelling on the Indie lyric collection, one of the main emerging topics was closely related to Reflection on life and the search for true love. Topic 0, or the first topic with *the song "babe, something tragic, something magic, agree babe, something lonesome, something wholesome..."* reflects the duality of emotions in a relationship—between bitterness and the beauty of love. The phrase *"familiar like a mirror from years ago"* gives a nostalgic feel to the love that was once known. Semantically, this topic raises the inner struggle to find true love's meaning amid doubt, irony, and emotional wounds, making love a personal emotion and an existential journey full of complexity.

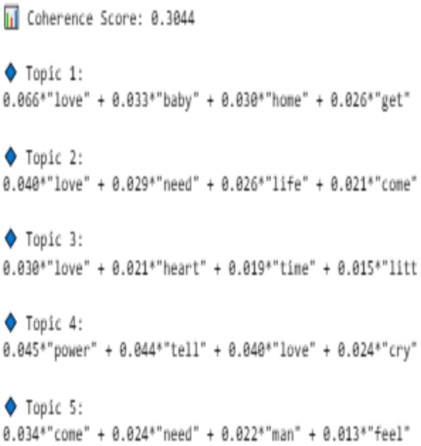


```
===== Topik 0 =====
🎵 babe something tragic something magic agree babe something lonesome someth
🎵 know know gonna collide know call end always always lie wear heart sleeve
🎵 alright cry even dad sometimes wipe eyes tears remind alive alright die ca

===== Topik 1 =====
🎵 birch tree lost branch one day violent winter said grieving said feel noth
🎵 soft hands weight world making mark delicate touch pulling pulling pulling
🎵 lightning strikes inside chest keep night dream ways make understand pain

===== Topik 2 =====
🎵 oh thousand nights new york london different city every day yeah playin pl
🎵 day kyoto got bored temple looked around eleven band took speed train went
🎵 breathing black white let fool caught linen eyes let fool know went dear u

===== Topik 3 =====
🎵 breaking back heart land feet water softened sand could wanted part find h
🎵 sometimes think killer scared house even scared talking dahmer couch sleep
🎵 twice walked life left find want back man dresses like john lennon sings t

===== Topik 4 =====
🎵 gonna pick pieces build lego house things go wrong knock three words two m
🎵 born inside small town lost state mind learned sing inside lord house stop
🎵 club best place find lover bar go friends table shots drinking fast talk s
```

Figure 10. Lyric quotes for each topic

Although the LDA model successfully grouped lyrics on five topics, semantic analysis showed overlaps and ambiguity between several topics. This is a common phenomenon in the topic of modelling on free texts such as musical lyrics. This case is, for example, in the first topic, with the second and third topics having the word love in the song's lyrics.

The evaluation of the LDA model conducted on the indie song dataset shows representative results with a coherence value of 0.3044. The score in Figure 11 shows that the semantic relationship in the song lyrics is at a level worthy of initial exploration, although the high coherence value has not been achieved. If you look at the distribution of the lyrics, there are many words love, time, and love, so these words occupy an important position, indicating that the theme of love, time, and emotional needs are the central motifs in the lyrics of indie songs. For example, topics 1 and 2 have more prominent nuances of romance and longing. Topic 4, with the words cry, believe, and break, emphasizes strength, belief, and emotion. Although there are similarities in words between topics, differences in focus can be seen in the context and dominant word pairs that appear. Topic 3 highlights the nuances of soft feelings.



```
📊 Coherence Score: 0.3044

♦ Topic 1:
0.066*"love" + 0.033*"baby" + 0.030*"home" + 0.026*"get"

♦ Topic 2:
0.040*"love" + 0.029*"need" + 0.026*"life" + 0.021*"come"

♦ Topic 3:
0.030*"love" + 0.021*"heart" + 0.019*"time" + 0.015*"litt

♦ Topic 4:
0.045*"power" + 0.044*"tell" + 0.040*"love" + 0.024*"cry"

♦ Topic 5:
0.034*"come" + 0.024*"need" + 0.022*"man" + 0.013*"feel"
```

Figure 11. *Coherence Score LDA*

Topic 5 emphasizes Reflection and introspection with the words feel and think. LDA successfully maps semantic variations to express lyrics, noting that there needs to be improvement to achieve a sharper interpretation. Overall, this study can provide initial insight into research, especially in thematic research on music. The documentation results show recurring patterns in the lyrics, such as the themes of love, struggle, and identity struggle, which often appear in indie music works.

Figures 4 and 5 will present the themes identified through LDA analysis, including the frequency of occurrence and examples of relevant lyrics. This image will visually represent the identified themes, illustrating the relationship between various themes in indie music lyrics. The study results show many hidden themes in indie music lyrics that were previously unidentified, enriching our understanding of this genre. By using LDA topic modelling, this study successfully uncovers complex patterns in the lyrics, overcoming the limitations of previous analysis methods. Documentation showing recurring themes provides empirical evidence to support the hypothesis that indie music lyrics reflect important social issues. Thus, this study contributes significantly to the literature on music and lyric analysis, paving the way for further study of themes in music.

Another form of topic modelling is NMF. This study uses initialization with n_components=5, which means the algorithm will try to identify **five** main topics from the text data represented as a TF-IDF matrix (X_tfidf). The init='nndsvda' parameter provides stable and efficient initialization for matrix decomposition, while max_iter=1000 ensures that the algorithm has enough iterations to achieve convergence. The fit_transform process will factor the TF-IDF matrix into two non-negative matrices: one that represents the distribution of the document against the topic and another that shows the distribution of words against the topic. The result is nmf_topics, a numerical matrix that shows how strongly each document (in this case, the song lyrics) represents each topic.

Table 1. Comparison *Coherence Score*

Model	Coherence
LDA	0,3044
NMF	0,2666

As a comparison pada Tabel 1 to the LDA model, which resulted in a coherence score of 0.3040, the Non-negative Matrix Factorization (NMF) model gave slightly lower results with a score of 0.2666, indicating that the topic structure produced by NMF tended to be less coherent than LDA in the collection of Indie song lyrics analyzed. Semantically, NMF was still able to group lyrics based on themes, such as Reflection on life and the search for true love, but the word transitions and the interconnectedness between words in its topics felt less natural than the LDA results. This can be due to the deterministic nature of NMF, which emphasizes linear decomposition and makes it more suitable for explicit and repetitive themes. In contrast, probabilistic-based LDAs can capture the diversity of context and implicit nuances in lyrics, making them superior at extracting latent meanings and hidden semantic structures.

CONCLUSION

The conclusion of this research study consists of:

This study has applied the CRISP-DM methodology as a systematic framework in topic modelling using Latent Dirichlet Allocation (LDA) to the indie song lyrics dataset.

The lyric generation dataset containing thousands of words from 347 songs has been applied with business and data understanding. After that, data preparation was carried out, namely initial data text processing with case folding, tokenization, stopword removal, and word normalization. The preparation results were to analyze text data so that they focused more on the thematic meaning of the song lyrics.

Hidden themes have been successfully identified in the indie song lyrics, thereby increasing understanding of the meaning and context of this song.

WordCloud visualization on the indie dataset shows the dominant words, namely love, time, need, and life, which indicate the major themes in this dataset. After LDA modelling was carried out with several topics, WordCloud visualization showed a diversity of semantic focus between topics. In addition, the proportion of topics shows that most of the lyrics have several dominant topics, so it can be assumed that this indie music dataset is multi-dimensional.

Love and longing are dominant topics because they appear several times in many indie songs. Insight into several themes shows the emergence of topic distribution in various works in a spread manner.

The LDA model produces a Coherence Score of 0.3044, indicating decent semantic coherence, but it can still be improved with several improvements approaches and other models.

Using the CRISP-DM methodology, the topic modelling approach provides a meaningful thematic mapping of the indie dataset lyrics. This research can be the basis for further research on music recommendation systems, genre classification, and semantic research in musicology.

The data will be analyzed using the LDA topic modelling technique to group lyrics based on the identified themes. The analysis results will be presented in narrative and tabular form to support the research findings.

The study was limited to 347 row-long data for further research using a larger, multilingual dataset with the addition of validation by humans as well. Further research also needs a multilingual corpus to add paragraphs explicitly stating limitations (e.g., dataset size and lack of human validation).

REFERENCES

[1] S. Dua *et al.*, “Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network,” *Appl. Sci.*, vol. 12, no. 12, p. 6223, Jun. 2022, doi: 10.3390/app12126223.

[2] H. Luo *et al.*, “Human–Machine Interaction via Dual Modes of Voice and Gesture Enabled by Triboelectric Nanogenerator and Machine Learning,” *ACS Appl. Mater. Interfaces*, vol. 15, no. 13, pp. 17009–17018, Apr. 2023, doi: 10.1021/acsami.3c00566.

[3] B. Liu and Y. Lv, “The Influence of the Era of Big Data on Film and Television Art and Countermeasures,” no. Fmess, pp. 237–240, 2021.

[4] M. de Witte, A. Spruit, S. van Hooren, X. Moonen, and G.-J. Stams, “Effects of music interventions on stress-related outcomes: a systematic review and two meta-analyses,” *Health Psychol. Rev.*, vol. 14, no. 2, pp. 294–324, Apr. 2020, doi: 10.1080/17437199.2019.1627897.

[5] Y. R. Pandeya, B. Bhattarai, and J. Lee, “Deep-Learning-Based Multimodal

Emotion Classification for Music Videos," *Sensors*, vol. 21, no. 14, p. 4927, Jul. 2021, doi: 10.3390/s21144927.

[6] Y. R. Pandeya and J. Lee, "Deep learning-based late fusion of multimodal information for emotion classification of music video," *Multimed. Tools Appl.*, vol. 80, no. 2, pp. 2887–2905, Jan. 2021, doi: 10.1007/s11042-020-08836-3.

[7] M. de Witte, A. da S. Pinho, G. Stams, X. Moonen, A. E. R. Bos, and S. van Hooren, "Music therapy for stress reduction: a systematic review and meta-analysis," *Health Psychol. Rev.*, vol. 16, no. 1, pp. 134–159, Jan. 2022, doi: 10.1080/17437199.2020.1846580.

[8] H. Richard, P. Dornheim, and T. Weber, "Using AI to Improve Risk Management : A Case Study of a Leading Telecommunications Provider," *IEEE Access*, vol. 12, no. November, pp. 165068–165080, 2024, doi: 10.1109/ACCESS.2024.3488321.

[9] H. Mamdouh and F. Tarek, "A high-quality feature selection method based on frequent and correlated items for text classification," *Soft Comput.*, vol. 27, no. 16, pp. 11259–11274, 2023, doi: 10.1007/s00500-023-08587-x.

[10] N. Jalal, A. Mehmood, G. Sang, and I. Ashraf, "A novel improved random forest for text classification using feature ranking and optimal number of trees," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 2733–2742, 2022, doi: 10.1016/j.jksuci.2022.03.012.

[11] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train , Prompt , and Predict : A Systematic Survey," vol. 55, no. 9, 2023, doi: 10.1145/3560815.

[12] T. H. Tee, B. Q. Bei Yeap, K. H. Gan, and T. P. Tan, "Learning to Automatically Generating Genre-Specific Song Lyrics: A Comparative Study," 2022, pp. 62–75. doi: 10.1007/978-3-031-21422-6_5.

[13] M. Mayerl, S. Brandl, G. Specht, M. Schedl, and E. Zangerle, "Verse Versus Chorus: Structure-Aware Feature Extraction for Lyrics-Based Genre Recognition," *Proc. 23rd Int. Soc. Music Inf. Retr. Conf. ISMIR 2022*, pp. 884–890, 2022.

[14] S. Sharma, A. Shukla, A. Walimbe, T. Sharma, and J. Delgado, "LyricLure: Mining Catchy Hooks in Song Lyrics to Enhance Music Discovery and Recommendation," in *18th ACM Conference on Recommender Systems*, Oct. 2024, pp. 800–802. doi: 10.1145/3640457.3688049.

[15] I. Czedik-Eysenberg, O. Wieczorek, A. Flexer, and C. Reuter, "Charting the Universe of Metal Music Lyrics and Analyzing Their Relation to Perceived Audio Hardness," *Trans. Int. Soc. Music Inf. Retr.*, vol. 7, no. 1, Aug. 2024, doi: 10.5334/tismir.182.

[16] A. M. Demetriou, J. Kim, S. Manolios, C. C. S. Liem, and S. Pandora, "TOWARDS AUTOMATED ESTIMATION OF VALUES FROM SONG LYRICS : A DATA COLLECTION PROTOCOL," pp. 57–59, 2023.

[17] A. Lukic, "A Comparison of Topic Modeling Approaches for a Comprehensive Corpus of Song Lyrics Dataset Methodology," pp. 1–7.

[18] S. Rani, "SENTIMENT ANALYSIS AND TOPIC MODELLING ON TWITTER FOR CLEAN INDIA MISSION," vol. 12, no. 5, pp. 1198–1207, 2021.

[19] W. Chen, F. Cai, H. Chen, and M. D. E. Rijke, "Personalized query suggestion diversification in information retrieval," vol. 14, no. 3, 2020.

[20] E. K. Seltzer *et al.*, "Patient Experience and Satisfaction in Online Reviews of Obstetric Care : Observational Study Corresponding Author :," vol. 6, pp. 1–8, doi: 10.2196/28379.

[21] K. Siriket, V. Sa-ing, and S. Khonthapagdee, "Mood classification from Song Lyric using Machine Learning," in *2021 9th International Electrical Engineering Congress (iEECON)*, Mar. 2021, pp. 476–478. doi: 10.1109/iEECON51072.2021.9440333.

[22] M. D. Devi and N. Saharia, "Exploiting Topic Modelling to Classify Sentiment from Lyrics," 2020, pp. 411–423. doi: 10.1007/978-981-15-6318-8_34.

[23] D. Yang, X. Chen, and Y. Zhao, "A LDA-Based Approach to Lyric Emotion Regression," 2011, pp. 331–340. doi: 10.1007/978-3-642-25661-5_43.

[24] T. Hunke, F. Huber, and J. Steffens, “The Evolution of Song Lyrics: An NLP-Based Analysis of Popular Music in Germany from 1954 to 2022,” *Music Sci.*, vol. 8, Apr. 2025, doi: 10.1177/20592043251331155.

[25] S. Zhang, R. C. Repetto, and X. Serra, “Understanding the expressive functions of jingju metrical patterns through lyrics text mining,” *Proc. 18th Int. Soc. Music Inf. Retr. Conf. ISMIR 2017*, pp. 397–403, 2017.

[26] P. Kherwa and P. Bansal, “A Comparative Empirical Evaluation of Topic Modeling Techniques,” 2021, pp. 289–297. doi: 10.1007/978-981-15-5148-2_26.

[27] J. BRZOZOWSKA, J. PIZOŃ, G. BAYTIKENOVA, A. GOLA, A. ZAKIMOVA, and K. PIOTROWSKA, “DATA ENGINEERING IN CRISP-DM PROCESS PRODUCTION DATA – CASE STUDY,” *Appl. Comput. Sci.*, vol. 19, no. 3, pp. 83–95, Sep. 2023, doi: 10.35784/acs-2023-26.

[28] O. Azeroual, R. Nacheva, A. Nikiforova, and U. Störl, “A CRISP-DM and Predictive Analytics Framework for Enhanced Decision-Making in Research Information Management Systems,” vol. 49, pp. 67–86, 2025.

[29] J. Bokrantz, M. Subramaniyan, and A. Skoogh, “The Management of Operations Realizing the promises of artificial intelligence in manufacturing by enhancing CRISP-DM,” *Prod. Plan. Control*, vol. 35, no. 16, pp. 2234–2254, 2024, doi: 10.1080/09537287.2023.2234882.

[30] I. Kolyshkina and S. Simoff, “Interpretability of Machine Learning Solutions in Public Healthcare : The CRISP-ML Approach,” vol. 4, no. May, 2021, doi: 10.3389/fdata.2021.660206.

[31] M. Konrad and X. State, “Automatic Complaints Classification in E-Commerce : A Case Study Using CRISP-DM,” 2025, doi: 10.5753/jis.2025.4661.

[32] C. Schröer, F. Kruse, and J. M. Gómez, “A Systematic Literature Review on Applying CRISP-DM Process Model,” *Procedia Comput. Sci.*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[33] C. Schröer, F. Kruse, J. Marx, F. Kruse, and J. Marx, “ScienceDirect ScienceDirect A Systematic Literature Review A Systematic Literature Review on Applying Process Model on Applying CRISP-DM Process Model,” *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[34] N. Cavus, M. Goksu, and B. Oktekin, “Real-time fake news detection in online social networks : FANDC Cloud-based system,” pp. 1–11, 2024.

[35] R. Nisbet, K. McCormick, and G. Miner, “Data Understanding,” in *Handbook of Statistical Analysis*, Elsevier, 2025, pp. 69–74. doi: 10.1016/B978-0-443-15845-2.00006-2.

[36] Z. Ma and B. N. Jørgensen, “DataPro – A Standardized Data Understanding and Processing Procedure : A Case Study of an Eco-driving Project,” pp. 1–20.