Topic Modeling on Taylor Swift Songs

Ayusee Swain, Khatina Sari

Introduction

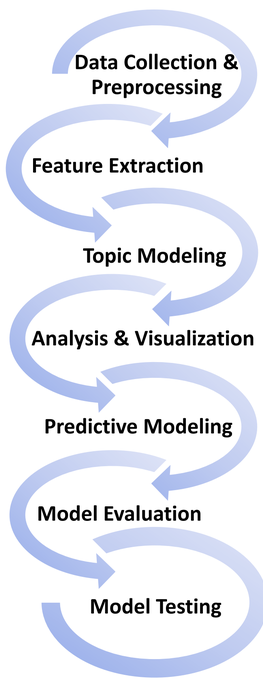
This project explores the themes in Taylor Swift's lyrics using topic modeling and predictive modeling techniques. We aim to uncover the prominent themes in her songs, analyze how these themes vary across different albums, and predict album, release year, or thematic category based on the lyrics. We can achieve this by utilizing unsupervised machine learning algorithms for topic modeling, such as LDA and NMF. These techniques cluster words that frequently appear together across multiple documents.

Latent Dirichlet Allocation (LDA) is a generative probabilistic model that extracts hidden topics from a given corpus by iteratively estimating word distributions across topics and topic distributions across documents, enabling the identification of distinct themes within the text [1].

Non-negative Matrix Factorization (NMF) is an algebraic model that factorizes non-negative matrix (X) into the product of two lower-dimensional matrices (A and B), so that AB approximately represents an optimal solution to X [2]. This reduces the dimensionality of data into lower-dimensional spaces.

Once the features are extracted, combined with other additional features, and the data dimension is reduced, we train our predictive models using **Random Forest** and **Support Vector Machine**. Following that, in order to determine which of these methods performs better at predicting the album, year, and theme of any given lyrics, we examine the accuracy, prediction, recall, and f1 score of each technique.

Methods



The first step in the project is **data collecting**, where we compile a list of Taylor Swift's albums, songs, and lyrics. Next, we proceed with **data preprocessing** by eliminating any irrelevant information. To standardize the text and enable more accurate and efficient data analysis, we separate the lyrics into individual words or tokens (tokenization) and reduce them to their base or root form (lemmatization).

The next step is **feature extraction**, which involves two methods: first, we use Count Vectorizer to convert text to a matrix of token counts that show the frequency of each word in the document; second, we use TF-IDF to adjust word frequencies based on the importance of words across the entire dataset, giving common words less weight [3].

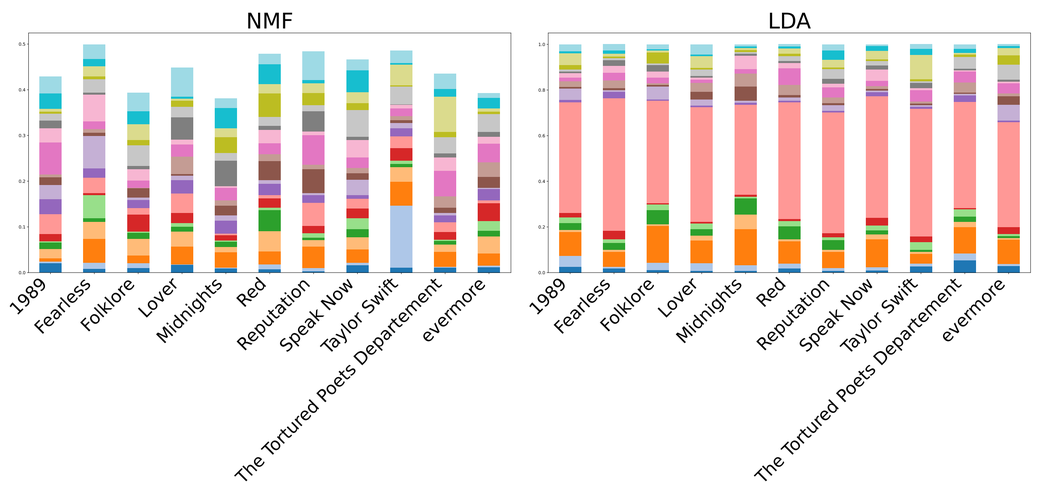
Following feature extraction, **topic modeling** is carried out using LDA and NMF. Afterwards, we **analyze and visualize** the top keywords for every topic, the document distribution across topics for both techniques and the theme variation across different albums. We also included the word clouds visualization for every album separately and as a whole to provide a comprehensive overview of the most prominent words used across all albums in the dataset. Larger words in the word cloud represent those that appear more frequently in the lyrics.

Using the features extracted by NMF, we apply SMOTE to balance the classes, followed by PCA to reduce the dimensionality of the feature matrices in the **predictive modeling** phase. The balanced dataset is then split into training and testing sets using stratification before being trained using Random Forest and Support Vector Machine (SVM). Subsequently, we **evaluate the model's performance** using standard metrics (recall, accuracy, precision, and F1 score). By entering a few lines of lyrics in the final phase, **model testing**, we can obtain album, year, and theme category predictions.

Model Comparison

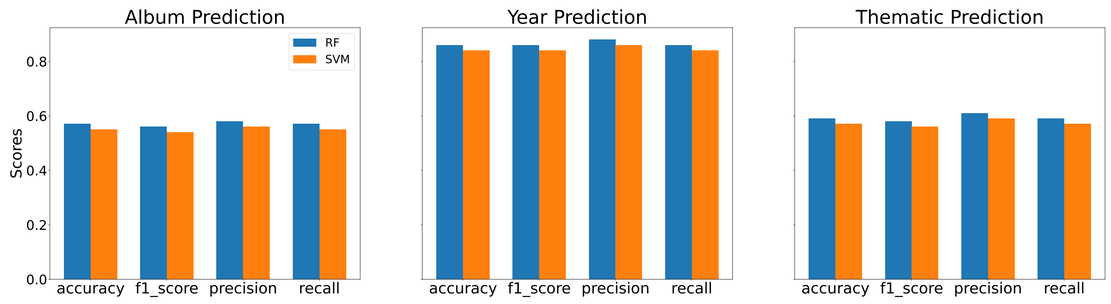
NMF vs LDA

The following bar chart illustrates the variation of 20 themes across different albums using NMF and LDA. Each bar represents the average topic distribution for an album, with different colors indicating distinct themes identified from the lyrics. NMF provides a more balanced topic distribution compared to LDA, which results in a single topic dominating.

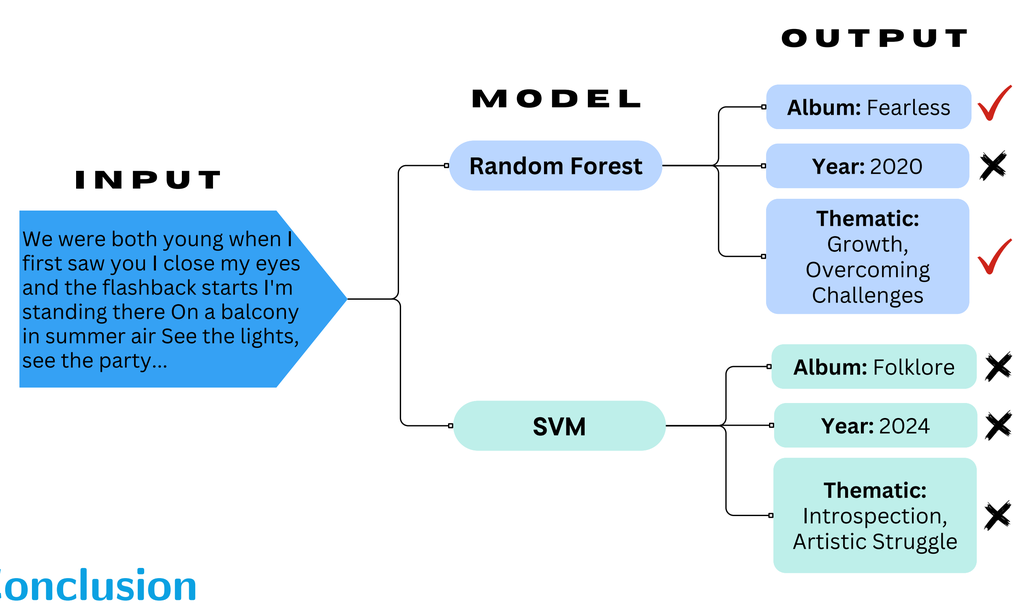


Random Forest vs SVM

Our predictive modeling comparison reveals that Random Forest scores slightly better on the evaluation metrics as compared to SVM.



However, the true distinction between these two models becomes evident during testing. Random Forest is able to accurately identify the album and theme when given multiple lyrics, whereas SVM not only predicts wrongly but also appears to produce the same result for two different inputs.



Conclusion

This project combines topic and predictive modeling to create pipelines for text analysis and prediction tasks. We can see that NMF produces better results in topic modeling compared to LDA. In predictive modelling, Random Forest outperforms SVM since RF is an ensemble method combining multiple decision trees, capturing various patterns and non-linear relationships, making it robust to overfitting, whereas SVM is effective in high-dimensional spaces and might struggle with subtle patterns, potentially leading to similar predictions for related inputs. These findings demonstrate the potential and constraints of using NLP methods to cultural and artistic analysis. To increase model performance, future developments could focus on incorporating advanced data processing techniques.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Sajad Fathi Hafshejani and Zahra Moaberfard. Initialization for non-negative matrix factorization: a comprehensive review. *International Journal of Data Science and Analytics*, 16(1):119–134, 2023.

[3] Sebastian Raschka. Musicmood: Predicting the mood of music from song lyrics using machine learning. *arXiv preprint arXiv:1611.00138*, 2016.

Acknowledgement

Special thanks to Prof. Dr. Jörn Hees and M.Sc. Tim Metzler for their invaluable guidance and support throughout this project. We extend our heartfelt gratitude to Taylor Swift for her prolific songwriting, which provided the rich dataset essential for our analysis.

Contact

Ayusee Swain/Khatina Sari
Hochschule Bonn-Rhein-Sieg
Email: ayusee.swain/khatina.sari@smail.inf.h-brs.de

b-it Bonn-Aachen International Center for Information Technology

Hochschule Bonn-Rhein-Sieg
University of Applied Sciences