NLP USING NMF

Twittes

Wadiqa Baig

August 30, 2025

Contents

1 Introduction

Topic modelling and Natural Language Processing (NLP) are two powerful ways of processing unstructured data in social media. In this report, these methods will use the Netflix series, Squid Game, that gained remarkable worldwide popularity in 2021 [1]. The viral success of the show was not necessarily the result of the storyline or the production. However, the show gained more publicity due to the interactions on social media, particularly on Twitter [2]. Using Non-negative Matrix Factorisation (NMF) and Term Frequency-Inverse Document Frequency (TF-IDF) vectorisation, this report identifies central themes in tweets related to Squid Game, offering insights into how social media influences cultural trends and how people respond to them.

2 Data collection and pre-processing

2.1 Data Source Description

The dataset consists of user-generated tweets about the famous Netflix show Squid Game and demonstrates a variety of reactions, opinions, and themes expressed by Twitter users because of the show's popularity [3]. This dataset was considered appropriate due to its cultural relevance and the high volume of user interaction it reflects.

2.2 Justification for Data Selection

To maintain computational efficiency and minimise memory use in Google Colab, 80,019 tweets were trimmed down to 30,000 tweets. A random sample was chosen to keep the contents diverse and enhance pre-processing and modelling efficiency. Even after reducing the tweets were a large enough and diverse source of text that could be used to train an NMF model. It also reduced the tuning and visualisation stages and ensured faster convergence.

2.3 Data Cleaning Steps

A structural pre-processing operation was performed to analyse the tweets before doing the textural pre-processing. There were no retweets and null values, so the number of tweets remained the same. For topic modelling, only the column containing the tweet text was used. The most important textual cleaning processes are the following:

- **Removing URLs:** The web references were removed to exclude other irrelevant links.
- **Removing hashtags and mentions:** Hashtags, such as #SquidGame and mentions, such as @Netflix, were removed to concentrate on real content.
- **Lowercasing and Eliminating Special Characters:** A conversion to lowercase and removing punctuation/symbols was done for consistency.
- **Delete Numbers and Emoticons:** Digits and emoticons were deleted to focus on meaningful words.
- **Removal of stopwords:** Standard English stopwords (e.g., the, is, and) were removed via the NLTK stopword list to assign greater meaning to the terms.
- **Tokenisation and Whitespace Clean-up:** The texts were partitioned into tokens (words), and additional spaces were removed to regularise formatting.

These tweets, with a clean-up, were finally turned into TF-IDF vectors and provided to the NMF model to retrieve topics. Pre-processing was necessary to prevent the model from learning low-quality, uninterpretable topics on the remaining dataset.

3 Topic modelling building and explanation

3.1 Non-negative Matrix Factorisation (NMF)

In contrast to other techniques, NMF is a factorisation algorithm that decomposes high-dimensional data structures into lower-dimensional structures that can be understood or have a meaning [4]. TF-IDF vectorised textual data is input to the algorithm, which factorises the original document-term matrix (A) such that $A \approx W \times H$, where W and H are two matrices of reduced rank, comprising document-topics (W) and topic-terms (H) respectively [5].

The choice of NMF instead of the other potential methods, like LDA, was primarily because NMF is deterministic and has better performance testing with short-form textual data, which requires a well-defined topic to facilitate an effective analysis [6].

3.2 Implementation Details

The cleaned tweets were vectorised using TF-IDF to translate the texts into a numerical format. The TF-IDF technique was selected instead of count vectorisation since it can downweight frequent words and focus on those specific to documents.

The optimal hyperparameters of the TF-IDF vectoriser and the NMF model were found through a **grid search**.

TF-IDF Hyperparameters:

- `ngram_range = (1, 2)` to capture both unigrams and bigrams,
- `sublinear_tf = True` for logarithmic term frequency scaling,
- `max_df` and `min_df` values were tuned to control how frequently terms appear across documents.

NMF was applied through `sklearn.decomposition.NMF`, where the data matrix is decomposed into two lower-dimensional matrices, W and H.

NMF Hyperparameters:

- `n_components`**:** number of topics tested in the range of 8, 9, 20, and 30.
- `l1_ratio`**:** controlling sparsity, tested with values 0.2, 0.45, and 0.50.

The optimisation procedure indicated that the optimum setting contained 8 topics with `l1_ratio` of 0.45, `max_df` of 0.5 and `min_df` of 1 and had the best score of **0.768** coherence in the validation argument.

Tools and Libraries Used:

- `scikit-learn` for TF-IDF and NMF modelling,
- `gensim` for computing coherence scores,
- `matplotlib, seaborn, pyLDAvis` for visualisation,
- `NLTK` for stopword removal and stemming,
- `wordcloud` to illustrate the most representative words per topic.

4 Unsupervised topic modelling evaluation

The most used evaluation metric was the `c_v` coherence score, which reports the degree of semantic coherence of topics by calculating co-occurrence patterns of top words in each topic [7]. The final model reported a coherence score of 0.**7684** (76.84%), which indicates high topic quality and interpretability.

A unique trend appeared when optimising the size of the dataset and found that, at the reduction to 20,000 tweets, the coherence score was at 71%, which was slightly increased when the dataset was increased to 30,000 tweets, leading to a score of 76% but with further expansion (up to 40,000 tweets), the score declined to 68%. This suggests an optimal balance between dataset size and topic coherence.

5 Topic modelling results, visualisation, and interpretation

The pyLDAvis visualisation (Figure 1 and Figure 2) offers vital information concerning the topic relationship and quality. The intertopic distance map shows a well-spaced topic cluster with few overlaps; thus, the topic differentiation is successful. Topics are shown as separate circles in the two-dimensional space, and the larger the circle is, the higher the topic's prevalence in the corpus.

The distance between topic centres in the visualisation is related to the semantic similarity, such that closely positioned topics have a higher semantic content.

Figure 1: Intertopic Distance Map of Topic 1

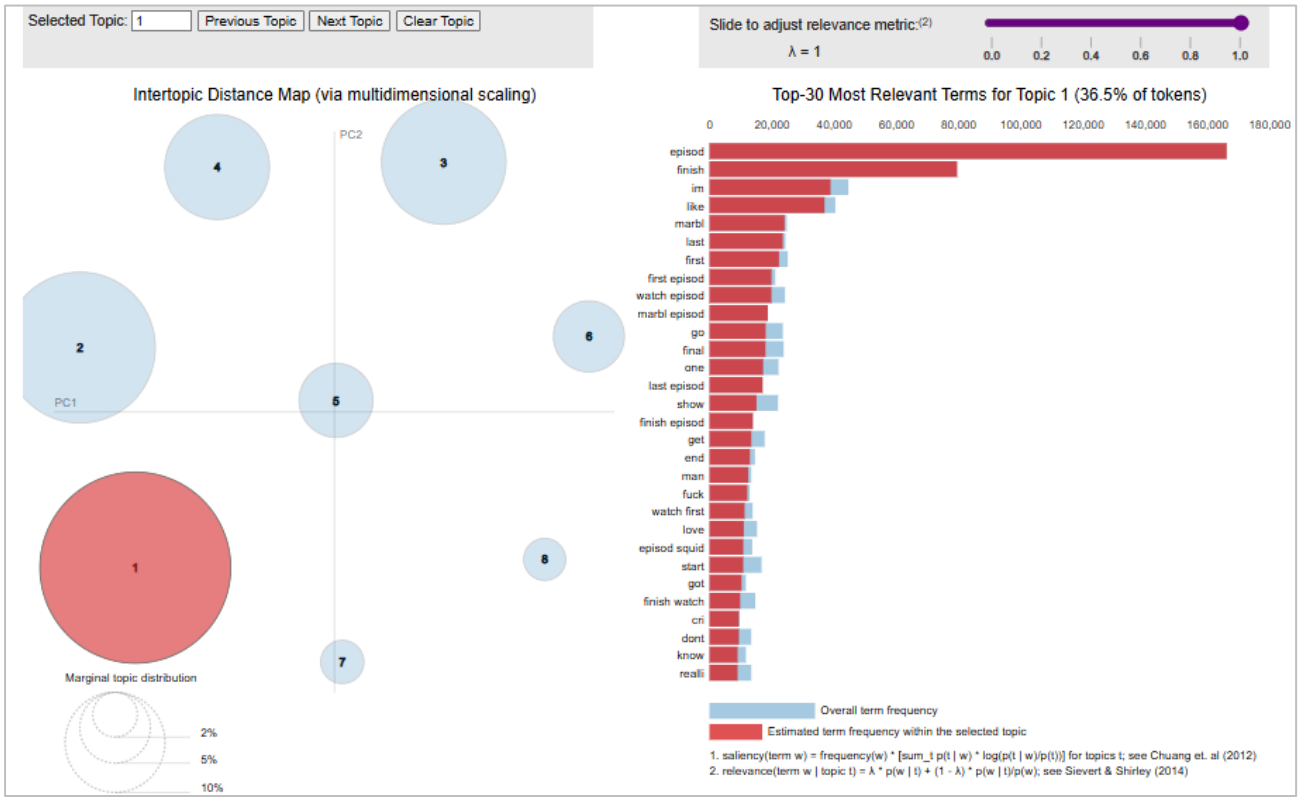
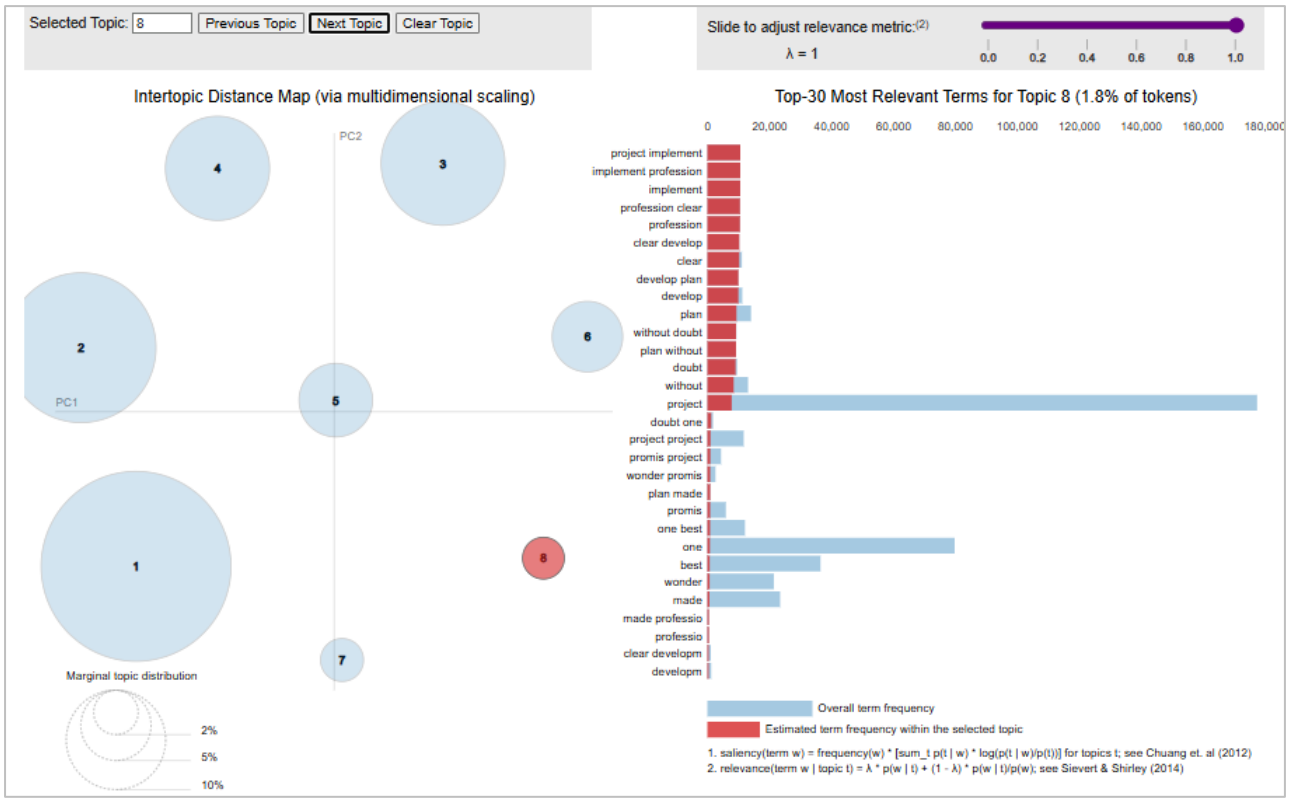
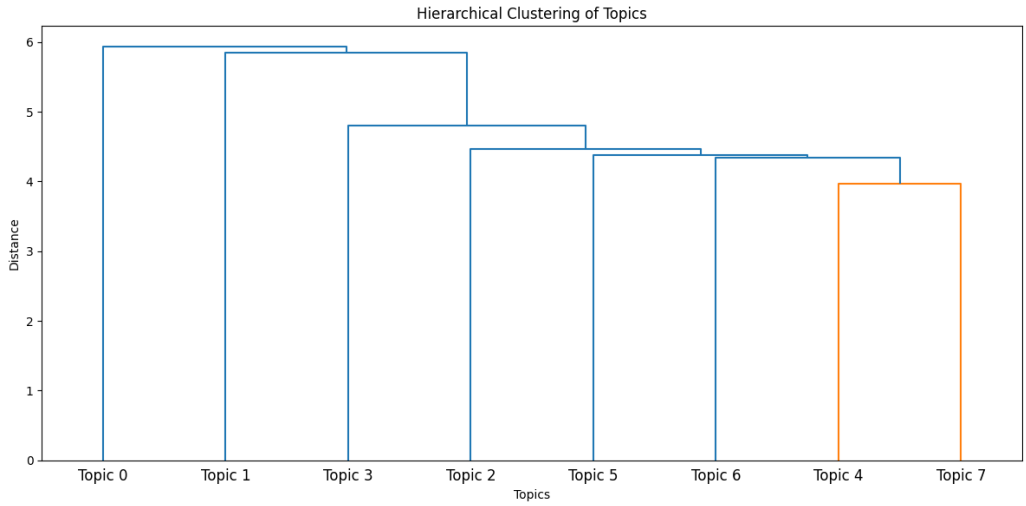


Figure 2: Intertopic Distance Map of Topic 8



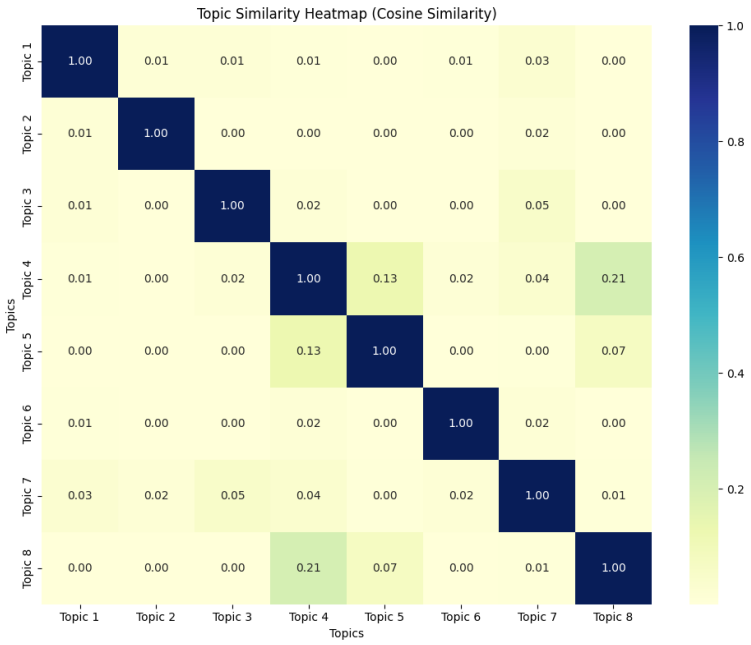
The Figure 3: Hierarchical Clustering of Topics shows conceptual groupings between the topics, which implies the thematic clusters that may be the higher-order categories of discussion about the Squid Game.

Figure 3: Hierarchical Clustering of Topics



The Figure 4: Topic Similarity Heatmap (Cosine Similarity) shows that topics are distinctive and topics 4 and 8 are moderately similar (values in the 0.21), which means related but slightly different topics. This degree of similarity is ideal for topic modelling since it suggests that topics are neither completely isolated nor excessively overlapping.

Figure 4: Topic Similarity Heatmap (Cosine Similarity)



Term distributions of the bar chart visualisations (Figure 5) display attention to the clear distributions of the terms in each topic, and there are pronounced peaks that reveal the most characteristic vocabulary of each theme. The TF-IDF values vary significantly, proving that the model identified unique term patterns instead of generic vocabulary.

Figure 5: Plot Top Terms per Topic (Bar Chart)

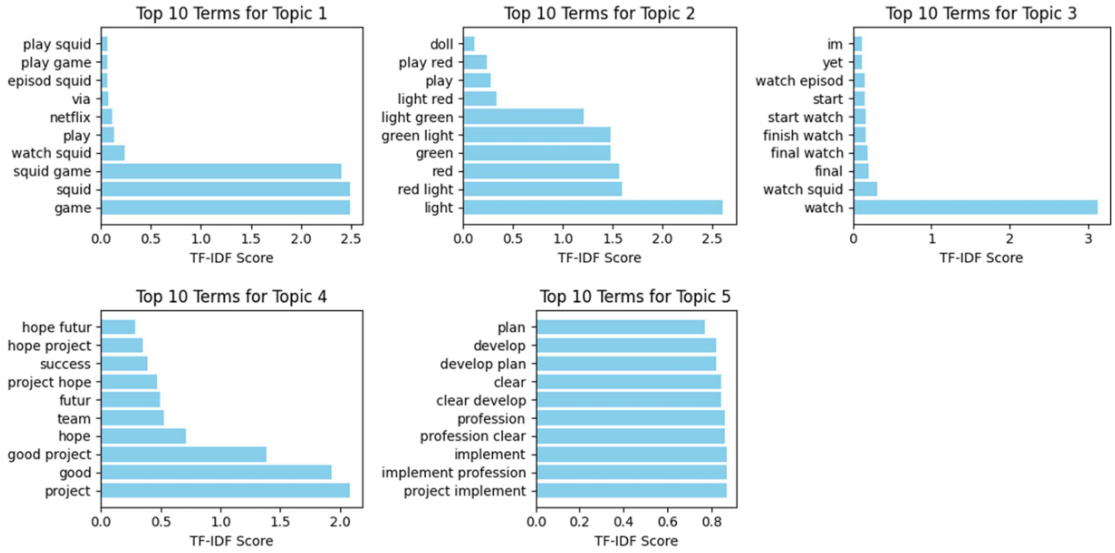


Figure 6 shows the Word cloud representations of all 8 topics. These are visually recognised as the most frequently used and influential words for each topic, with the larger fonts having a higher relevance.

Figure 6: Word Clouds for Topics (Topic 1-8)

6 Conclusion

This report has demonstrated that NMF could be used to extract topics from Twitter conversations about the Netflix series Squid Game. The NMF model produced a 76.84% coherence score with each topic identified only by a small quantity of overlap, which is effectively 8 different topics.

Although the given analysis studied only the tweets in English, the outcomes make a potential bias for automated topic discovery in social media discourse on cultural phenomena. Work in the future would be an extension of the current approach through temporality, multi-language implementation, and cross-platform comparison to give a wider picture of digital cultural discourse patterns.

7 References

[1] M. M. Nan, "*Squid Game*: The Hall of Screens in the Age of Platform Cosmopolitanism," *Global storytelling*, vol. 3, no. 1, Jul. 2023, doi: https://doi.org/10.3998/gs.4156

[2] W. Ahmed, A. Fenton, M. Hardey, and R. Das, "Binge Watching and the Role of Social Media Virality towards promoting Netflix's Squid Game," *IIM Kozhikode Society & Management Review*, vol. 11, no. 2, p. 227797522210833, Mar. 2022, doi: https://doi.org/10.1177/22779752221083351

[3] D. Contractor, "Squid Game Netflix Twitter Data," Kaggle.com, 2021. Available: https://www.kaggle.com/datasets/deepcontractor/squid-game-netflix-twitter-data/data

[4] K. Chen, J. Liang, J. Liu, W. Shen, Z. Xu, and Z. Yao, "Entropy regularized fuzzy nonnegative matrix factorization for data clustering," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 2, pp. 459–476, Jul. 2023, doi: https://doi.org/10.1007/s13042-023-01919-1

[5] A. Yavari, H. Hassanpour, B. Rahimpour Cami, and M. Mahdavi, "Event prediction in social network through Twitter messages analysis," *Social Network Analysis and Mining*, vol. 12, no. 1, Jul. 2022, doi: https://doi.org/10.1007/s13278-022-00911-x

[6] S. Si, J. Wang, R. Zhang, Q. Su, and J. Xiao, "Federated Non-negative Matrix Factorization for Short Texts Topic Modeling with Mutual Information," *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, Jul. 2022, doi: https://doi.org/10.1109/ijcnn55064.2022.9892602

[7] A. Simonetti, A. Albano, A. Plaia, and M. Tumminello, "Ranking coherence in topic models using statistically validated networks," *Journal of Information Science*, vol. 51, no. 3, Jan. 2023, doi: https://doi.org/10.1177/01655515221148369