Stand-Up Comedy Topic Modeling Across Cultures

Lonny Chen (216697)
Research Note for Natural Language Processing (E1282)
Dr. Sascha Göbel, Fall 2025

8 January 2026

1 Abstract

This research aims to use the Latent Dirichlet Allocation (LDiA) method of topic modeling to explore and compare differences in topics used for stand-up comedy across cultures. The dataset is collected from transcripts fetched from YouTube Shorts from Comedy Central's global channels. The study is currently limited to English-language channels catering to US and UK audiences. The methodology faithfully follows a structured Natural Language Processing (NLP) pipeline of tasks, including iterative token selection and hyperparameter tuning. Five subjectively interpretable topics were found out of the total of 12 topics, and their distribution as the dominant topic of the Shorts is largely similar between the US and UK channels. The code and data are available online at https://github.com/lonnychen/nlp_research_note.

2 Introduction

Humour and comedy are essential for social interaction, and their recurrent themes can be thought of as cultural markers that ebb and flow across immediate context, social identity, geography, and time. Comedy involves notions of counter-intuition, surprise, and absurdity. Thus, programming computers to understand comedy has been challenging (Turano & Strapparava, 2022).

Stand-up comedy is a specific form of performance comedy that holds promise for deeper computational insights. Comedians prepare sequences of jokes that are often, but not always, grouped by meaningful topics. Topics and punchlines are tuned by comedians to align with the humorous preferences of specific audiences.

A comparison of popular topics used for stand-up comedy across cultures could be useful in a range of applications: cross-cultural understanding, tracking trends of humour and socio-economic opinions, and improving the humour embedded in large language models.

Topic modeling is a subfield of both Natural Language Processing (NLP) and machine learning. It is a specific technique of unsupervised machine learning as it aims to discover latent patterns or groupings in its input data, usually a corpus of documents. These patterns or groupings can often be used for further downstream applications.

This study aims to answer the research question of whether prominent differences in topics used by stand-up comedians from and for different cultures can be discovered by applying current topic modeling methods to a dataset of textual transcripts of their comedic material. It is largely inspired by Barriere et al. (2025)'s curation of the StandUp4AI dataset for multicultural and multilingual humour detection from Comedy Central's global YouTube channels.

3 Background

3.1 Latent Dirichlet Allocation Overview

A widely used method for topic modeling is the Latent Dirichlet Allocation (LDiA) developed by Blei et al. (2003). LDiA is popular in the social sciences due to the interpretability of its two outputs, which provide semantic differentiation of documents, defined as meaningful collections of text, into topical weights and of the topics themselves into their constituent words from the dataset vocabulary.

The first output is the per-document topic distribution, denoted as θ. This enables documents to be understood as a weighted mixture of topics, which is realistic in many contexts. The second output is the per-topic word distribution, which represents the identifying word makeup of each topic, and is denoted by ϕ. Human domain judgment is needed to determine suitable meanings of the topics discovered by the LDiA model and use them for further analysis or learning. The key hyperparameter to set is the k number of topics to discover. This value needs to be determined exogenously or tuned by comparing an evaluation metric such as coherence.

The coherence score is one of the popular metrics for topic modeling and is based on the semantic similarity within each topic. It has been proven to show high correlation with human topic ranking (Syed & Spruit, 2017). It is often used to compare topic modeling methods, and its latest version, denoted C_v, relies on the co-occurrences of the set of top words (Fu et al., 2019). A coherence score ranges in $[0, 1]$ and a score of between $0.4 - 0.6$ is typically seen in the literature (Fu et al., 2019; Syed & Spruit, 2017).

3.2 Latent Dirichlet Allocation Mechanism

The probabilistic mechanism of fitting an LDiA model to its input data is well described in recent literature, by Lane and Dyshel, 2025, and López Pérez and Llinás Solano, 2025, for example, and videographically by Serrano.Academy, 2020. The underlying concept is a "document-generating machine" that randomly generates documents made up of words drawn from topics. The random generation is driven by the two balanced and competing goals of the sparsity (focus on fewer) of topics within documents and that of words within topics DiMaggio et al., 2013.

The algorithm uses a Dirichlet distribution as the Bayesian prior of a multinomial distribution for both document-topic and topic-word distributions as estimated from its input dataset. The objective function is to maximize the posterior likelihood of the latent topics given the sparsity settings of α and β for each distribution.

4 Methodology

The research methodology followed the standard NLP pipeline of data collection, text preprocessing, modeling for text understanding, and production of outputs for analysis of results. The modeling step included key parameter tuning for topic modeling.

All computational steps were programmed in Python. Two YouTube information extraction libraries were used for data collection: *YT-DLP* and *YouTube Transcript API*. Then the standard NLP libraries of *SpaCy*, *scikit-learn*, and *gensim* were used at specific stages in the pipeline – the important classes are noted below in monospaced font. The *Numpy*, *Pandas*, and *Seaborn* libraries were used throughout the pipeline to support data manipulation and visualization.

4.1 Data Collection

Stand-up comedy is a fluid entertainment medium, and identifying well-defined topics for a comedian's show or set can be challenging even for humans, let alone an algorithmic text-based

model. Segmenting comedy topics at the "one joke" level was considered by using the setup-punchline-tags model described by Turano and Strapparava, 2022 and the laughter detection annotations generated by Barriere et al., 2025. However, the extra preprocessing steps and the inconsistent signaling characteristic of laughter (consider comedy segments with laughter buildup and multiple punchlines) were deemed beyond the scope of this study.

The chosen data source was YouTube Shorts of stand-up comedy. By definition, these short video segments have a duration of a maximum of 60 seconds, and the assumption is that they each feature a shorter comedy sequence of related jokes, also known as a "comedy bit". The targeted YouTube channels were the US-centric Comedy Central Stand-Up specialty channel and the Comedy Central UK channel, allowing a cultural comparison of comedic topics that were assumed to be curated for the American and British audiences, respectively.

Barriere et al., 2025's curation of their StandUp4AI dataset provided guidance for extracting collections of comedy, particularly from Comedy Central YouTube channels globally. They used the *YT-DLP* library to extract video metadata, particularly the video identifier, from YouTube pages. We used the Shorts page for Comedy Central Stand-Up and a queried search results page for Comedy Central UK. Some trial and error was needed to find a reasonable search string of "standup" (no space or hyphen) for Comedy Central UK, and these results were then filtered by each video's duration field. A total of 768 video identifiers and their metadata were collected.

The *YouTube Transcript API* library was used to fetch transcripts from the collection of video identifiers. The first issue faced was an IP ban by YouTube due to repeated transcript fetches in a small amount of time. This was solved by using the paid *Webshare* online service to rotate residential IP addresses to avoid the ban. The second issue was related to the transcripts themselves. In our dataset, four videos had their transcripts disabled, and two videos only returned transcripts in a language other than English (both were in Dutch). An example transcript is provided in Appendix A.

4.2 Text Preprocessing

Data cleaning consisted of straightforward video-level and transcript-level tasks. At the video level, 51 video identifiers were found to be duplicated – only the first occurrence of each was kept in the dataset. At the transcript level, the previously flagged six video identifiers without English transcripts were dropped, leaving 711 videos in the dataset for preprocessing and topic modeling.

The test text for the examples below is:

> Hello world [Music].\nLet's go to "N.Y.", for 'around' - $200!? [Applause].

An iterative approach to tokenization, together with vectorization, was taken to monitor the forming and filtering of tokens using the scikit-learn library's 'CountVectorizer' class, which wraps a preprocessor, a tokenizer, stop word removal, and vectorizer into a Bag of Words output matrix.

1. *Baseline of preprocessor only.* The custom preprocessor only lowercases the text and removes the context-specific annotations (applause, laughter, music, profanity). For the test text, the output is:

   > hello world .\nlet's go to "n.y.", for 'around' - $200!? .

   This is then tokenized by *SpaCy*'s `en_core_web_sm` tokenizer, resulting in a vocabulary of 7,132 unique tokens.

2. *Remove non-words using SpaCy Token attributes.* The custom tokenizer now drops whitespace, punctuation, and pure digits, which were initially kept by `en_core_web_sm`. For the test text, the output becomes:

Step No.	Tokenization Step	Vocabulary Size
1	Baseline of preprocessor only	7,132
2	Remove non-words using SpaCy Token attributes	7,005
3	Remove SpaCy English stop words	6,735
4	Remove conversational stop words	6,726
5	Use lemmatized tokens instead of raw text	5,440
6	Remove single-occurrence tokens	2,349
7	Remove common words from baseline LDiA results	2,312

Table 1: Summary of the iterative tokenization steps and resulting vocabulary sizes.

['hello', 'world', 'let', "'s", 'go', 'to', 'n.y', 'for', 'around']

This reduced the vocabulary to 7,005 tokens.

3. *Remove SpaCy English stop words.* Topic modeling benefits from the removal of words without contextual meaning, technically known as "stopwords" so that modeled topics are more differentiable and interpretable (Schofield et al., 2017). *SpaCy*'s list of 326 English stop words was used to match and remove contraction components (such as "you" and "'ll" from "you'll") that are split from *SpaCy*'s tokenization. This reduced the vocabulary to 6,735 tokens.

4. *Remove conversational stop words.* A few conversational utterances such as "ah", "oh", and "um" are included in the transcripts due to the content's conversational nature. They are removed in this step, reducing the vocabulary to 6,726 tokens.

5. *Use lemmatized tokens instead of raw text.* Lemmatization, the combining of words into their grammatical base forms, has been shown to benefit the readability of topic terms (Bitext, 2023), and is considered essential for LDiA in particular by López Pérez and Llinás Solano, 2025 in order to focus on core meanings. Using *SpaCy*'s built-in lemma tokens reduced the vocabulary to 5,440 tokens.

6. *Remove single-occurrence tokens.* Further word removal by document frequency is supported by `CountVectorizer`'s `min_df` and `max_df` parameters. The `min_df` parameter was set to two to remove any words that appear only in one transcript, a common-sense approach. This resulted in the removal of 3,091 tokens.

7. *Remove additional common words after reviewing baseline LDiA results.* The initial baseline attempt at topic modeling with LDiA (described further below) revealed that pervasive common words such as "like", "know", "I", "right", and "think" still remained in the dataset, dominating the model's structure of each topic's most representative 10 words. A decision was made to remove words that appear in more than 10% of all documents using the `max_df` parameter, resulting in the removal of the most common 37 tokens.

A summary of the iterative filtering of the token vocabulary is provided in Table 1. The final distribution of transcript length in tokens is shown in Figure B1.

Zipf's law as described in detail by Piantadosi, 2014 states that word frequency in natural corpora decays by a power-law proportional to rank such that the second most frequent word has $\frac{1}{2}$ of the frequency as the most frequent word, the third has $\frac{1}{3}$, and so on. Zipf's law leads to a heavy tail of rare words and a straight line relationship on a log-log plot of rank versus frequency. Our dataset indeed follows Zipf's law as shown in the plots in Figure 1.

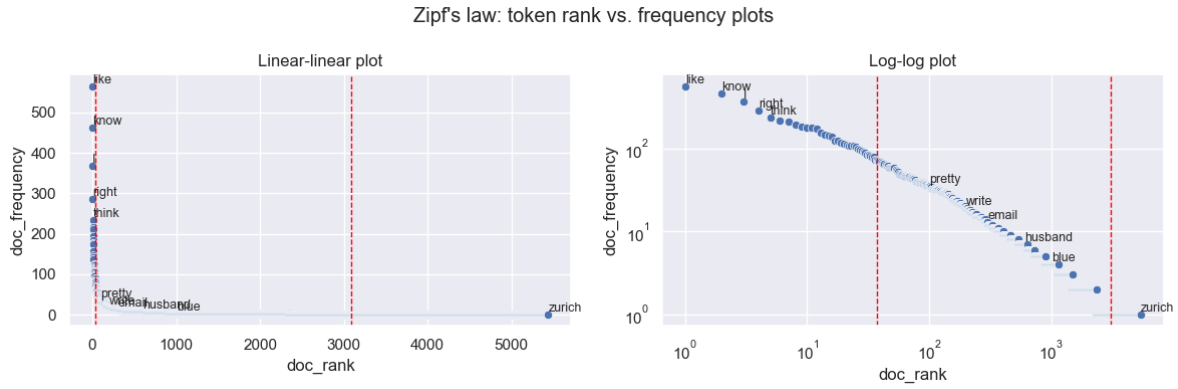

Figure 1: Plots of vocabulary tokens by document occurrence rank versus document occurrence frequency. The tokens between the two red lines are those kept in the final vocabulary based on the `min_df` and `max_df` parameters of `CountVectorizer`. A few example tokens are annotated for context.

4.3 Topic Modeling with Latent Dirichlet Allocation

Two approaches to the comparative topic modeling of US and UK stand-up comedy were considered. Fitting videos from both channels to one topic model was the chosen approach to enable comparison of relative emphasis on commonly found topics DiMaggio et al., 2013. The second considered approach of fitting separate models for each channel would have allowed for channel-specific discovery of topics and is included as a point for further work.

Scikit-learn's `LatentDirichletAllocation` class was used as the implementation of topic modeling with LDiA. One notable feature of this implementation is that it uses a variational Bayes algorithm to build its model instead of Gibbs sampling. Its direct output is a document-by-topic weight array that we ultimately use to compare the topic distribution between the two comedy channels. A post-modeling attribute is the topic-by-words array, which we processed to carefully review the top representative words for each topic for interpretation. In the LDiA literature, these two arrays correspond to the θ and ϕ distributions, respectively (López Pérez & Llinás Solano, 2025).

The baseline attempt at LDiA modeling used a k value of eight for the number of topics as input into `LatentDirichletAllocation`. This revealed that too many common words had still remained in the dataset, informing the last step of the Data Processing section: setting `min_df` to 10%. Further iterations of LDiA were run for hyperparameter tuning and analyzing the results of topic modeling using the best hyperparameters for LDiA, and are covered in the sections below.

Whereas k is tuned as described below, the α and β parameters, which control the distribution of topics per document and words per topic, respectively, use the model's default of $\frac{1}{k}$, which tends towards more topic-focused documents. The tuning of these two parameters is included as a point of future work.

4.4 Hyperparameter Tuning

The k number of topics is the key hyperparameter to tune for topic modeling and directly affects the structure and interpretability of the output (López Pérez & Llinás Solano, 2025; Syed & Spruit, 2017). Too few topics would make them too general (underfitting), and too many would make them too specific (overfitting) for interpretation and applicability. We did not employ cross-validation due to the already small size of the dataset.

Our tuning space consisted of five values for k between $[3, 15]$ and the coherence scores are presented in Table 2. The best k is at 12 topics with a coherence score of approximately 0.39.

k Topics	Coherence Score
3	0.2346
5	0.2791
8	0.3371
10	0.3259
12	0.3854
15	0.3727

Table 2: Hyperparameter tuning results using *gensim* `CoherenceModel`.

5 Results Analysis

Using the k setting of 12 topics from hyperparameter tuning, we proceeded to review the top 30 representative words for each topic. Per-topic plots of leading words' normalized distribution scores were especially helpful to know the extent to which they were more representative than other words. These are provided in Figure C1. Additional guidance was the response of the ChatGPT generative AI system to the prompt of *"Are these 12 topics interpretable?"* followed by the per-topic top words OpenAI, 2026.

For example, in topic_8, "dog" is weighted at 3% which is more than double that of the next top word, so this could be a more interpretable topic related to dogs. Similar logic is used to attempt to interpret other topics as described below in Table 3.

Topic	Top Word	Other related top words	Possible Themes
topic_8	"dog"	dog, cat, child, teacher, chase, play	Pets, children's leisurely activities.
topic_3	"gay"	gay, bless, date, muslim, marry, family	Gay culture and dating, family, religion
topic_12	"black"	black, white, kid, house, friend, street	Friendships between black and white people within a community.
topic_7	"dad"	dad, mom, girl, kid, home, family, boy	Parenting, family life
topic_9	"boy"	boy, date, party, girl, night, relationship	Youth dating, social life

Table 3: Subjectively interpretable topics from LDiA modeling.

The distributions of these interpretable topics, in terms of the dominant topic for each document, are separated by YouTube channel in Figure 2 to answer the key research question. The order and relative distributions of the five topics are similar between Comedy Central Stand-Up (US-centric) and Comedy Central UK. One discernible difference is that "black" topics are proportionally more prevalent on Comedy Central UK, whereas "dad" topics are proportionally more prevalent on Comedy Central Stand-Up.

A validation sample with one video from each topic was reviewed to confirm if the topic modeling corresponds to actual comedic content. The content was on-topic for four out of the five sampled videos. The off-topic video was from the "black" topic group, where the comedian used the word "black" to refer to "black out" or faint from consuming too much alcohol. This is an example of polysemy affecting the topic modeling results.

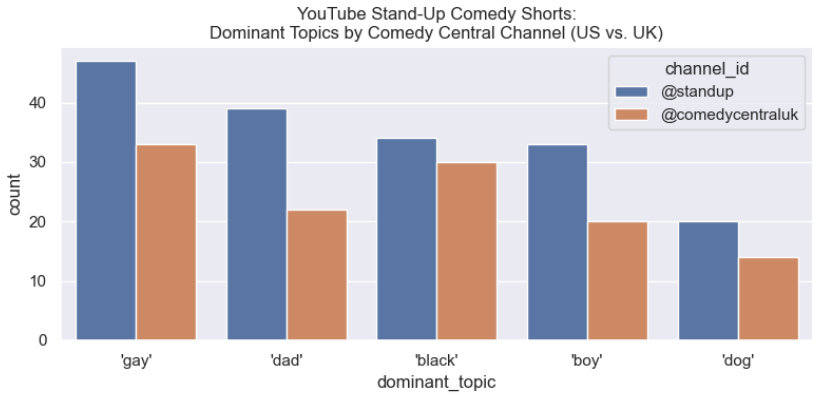

Figure 2: The dominant topic distribution is shown comparing the US (@standup) and UK (@comedycentraluk) channels. The two datasets are not exactly balanced and the counts are absolute values. Videos with "uninterpretable" topics were removed from this analysis.

6 Discussion and Conclusions

Prominent differences were not observed in the dominant topic distribution between the US-centric and UK-centric Comedy Centric YouTube channels for stand-up comedy YouTube Shorts. The fact that the distributions are similar may suggest some ability of LDiA to discover these five topics, among 12 total topics in the modeling, with focused and distinct comedic content, mostly related to daily life from various perspectives.

Spoken transcribed text may be more difficult to model compared to written text due to the nature of human speech and spoken performance. There are more filler words, off-topic asides, and unintended diction leading to generally less predictable topic separation than, for example, a written article that follows a defined structure and contextual vocabulary. Comedy itself may be especially challenging to model as it relies on unexpected, sometimes absurd, shifts in topic as the very structure of a joke.

The dataset itself may exhibit cross-pollution of the desired cultural comparison between the US and UK, especially because both channels are from the same global television network of Comedy Central. More and longer videos, and a greater variety of sources, may provide better representation.

Extensions for further work include:

- Fitting separate models for each channel to discover topics in each channel's content space.
- Tuning the α and β distribution parameters of the LDiA model.
- Extending this study to multilingual comedy content using Comedy Central's channels in other languages, an approach taken by Barriere et al., 2025.

References

Barriere, V., Gomez, N., Hemamou, L., Callejas, S., & Ravenet, B. (2025). Standup4ai: A new multilingual dataset for humor detection in stand-up comedy videos. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 16951–16959. https://doi.org/10.18653/v1/2025.findings-emnlp.919

Bitext. (2023). Lemmatization for topic modeling [Available online]. https://www.bitext.com/wp-content/uploads/2023/04/Topic-modeling.pdf

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent dirichlet allocation*. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

DiMaggio, P., Nag, M., & Blei, D. M. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, *41*(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Fu, Q., Zhuang, Y., Gu, J., Zhu, Y., Qin, H., & Guo, X. (2019). Search for k: Assessing five topic-modeling approaches to 120,000 canadian articles. *2019 IEEE International Conference on Big Data (Big Data)*, 3640–3647. https://doi.org/10.1109/BigData47090.2019.9006160

Lane, H., & Dyshel, M. (2025). *Natural language processing in action* (2nd). Manning Publications. https://www.manning.com/books/natural-language-processing-in-action-second-edition

López Pérez, C., & Llinás Solano, H. J. (2025). *A comprehensive guide to latent dirichlet allocation: Building a solid foundation with key statistical concepts* (tech. rep.). Universidad del Norte. https://doi.org/10.13140/RG.2.2.20062.34883

OpenAI. (2026, January). Chatgpt-5 response to the prompt "are these 12 topics interpretable?" [Accessed via https://chatgpt.com]. https://chatgpt.com

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

Schofield, A., Magnusson, M., & Mimno, D. M. (2017). Pulling out the stops: Rethinking stopword removal for topic models. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 432–436. https://doi.org/10.18653/v1/E17-2069

Serrano.Academy. (2020, March). Latent dirichlet allocation (part 1 of 2) [video] [Accessed 2025-12-31]. https://www.youtube.com/watch?v=T05t-SqKArY

Syed, S., & Spruit, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 165–174. https://doi.org/10.1109/DSAA.2017.61

Turano, B., & Strapparava, C. (2022). Making people laugh like a pro: Analysing humor through stand-up comedy. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5206–5211. https://aclanthology.org/2022.lrec-1.558/

A Transcript Example

An example transcript string fetched from https://www.youtube.com/shorts/JYjRBzuVFwQ (accessed on 4 January 2026) in a segment by comedian Chris Distefano is provided here:

> it was great being in a relationship\nwith her but I didn't know like Latino\nwomen like I didn't know that like\nyou're under investigation like it's\njust when you date a Puerto Rican girl\nyou need to have the answers quick you\nneed to like have facts right away\nbecause they will interrogate you and if\nyou don't have the answers quick you\nlook like a liar and they watch men they\nare designed to watch men watching\nwaiting that's what they do they make\nexcellent NFL referees I think if you\nhad Puerto Rican girls reing the games\nyou'd have zero Miss calls the whole\nseason because they see everything so\nthat's just what could you imagine they\nwere the refs they would be right there\nlike um you out of bound stupid no\nno I mean for real you out of\nbounds I don't need instant replay I I\ninstantly saw you step out of bound so\nyou\nknow I mean I don't even know the rules\njust go home\nbye you out the game

B Final Transcript Length Distribution

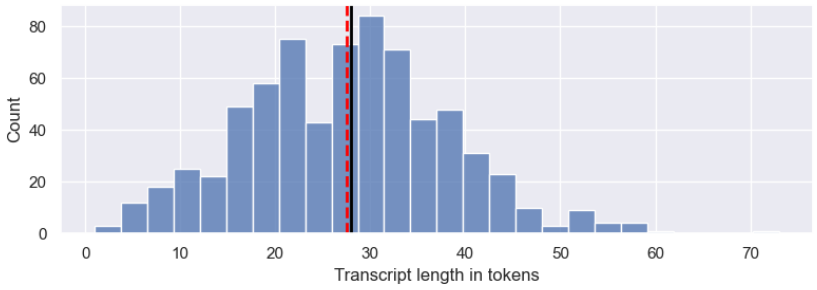

Figure B1: Transcript length distribution in tokens after all tokenization steps. The mean length is 27 tokens.

C Per-Topic Top 30 Words' Normalized Contributions

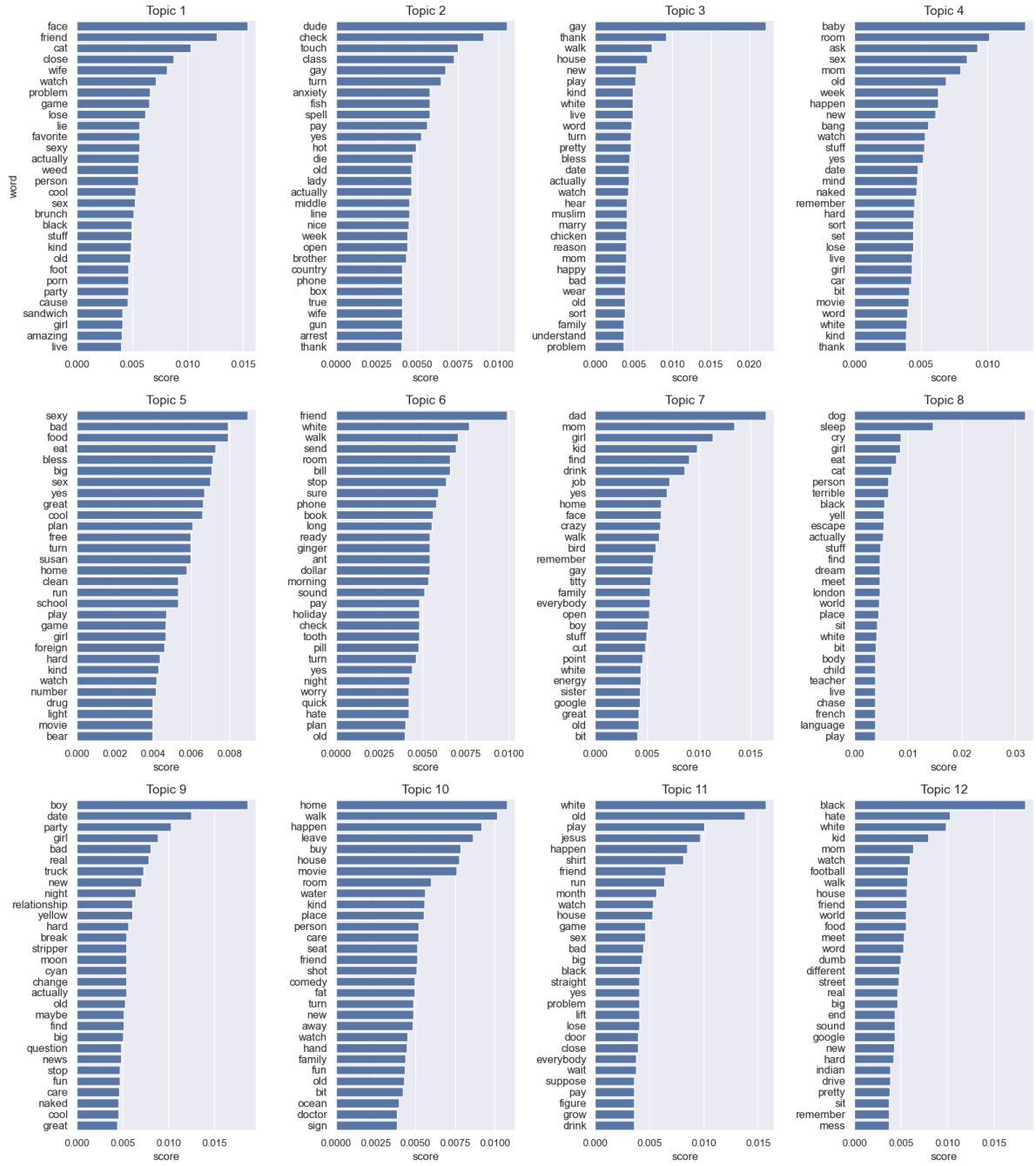

Figure C1: Normalized contribution plots of the top 30 words for each topic discovered by the LDiA model.