**Appendix B**

**Topic Model Creation**

The collection, preparation, and analyses of this data are summarized as follows:

1. Collection of the top 200 highest-grossing romcoms between 1980-2019 in the United States based on BoxOfficeMojo (n.d.).
2. Pre-processed scripts and prepared the text for analysis.
3. Scripts were split into scenes, considered the base unit of film, with a typical film comprised of 40-60 scenes (Murtagh, Ganz, & McKie, 2009). The average amount of scenes in our sample was 63 per film, ranging from 36 to 142.
4. Removed stopwords in multiple languages, from R’s quanteda dictionaries (e.g., “the”, “and”) and additional stopwords recommended for novels and literature, including first and last names from Jockers (2014), punctuations, and numbers.
5. Utilized ANTMN (Walter & Ophir, 2019) to calculate a topic model based on words’ frequency of co-occurrence across scenes. This creates a network wherein each node is a topic and the edges between nodes represent their co-occurrence in scenes.
6. Latent Dirichlet Allocation and Gibbs sampling (see Blei et al., 2003) grouped words into “topics” as statistical entities.
   1. Utilized a community detection algorithm (Fastgreedy; Clauset et al., 2004) to identify broader themes, operationalized as clusters of topics co-occurring in scenes.
   2. Utilized a “bag-of-words” approach which disregards narrative, location, and syntax are disregarded (Blei et al., 2003). Used 10-fold cross-validation comparing the perplexity scores of multiple models (see Walter, Ophir, & Jamieson, 2020), ranging between 10 and 120 topics. Analyses indicated optimal values in a model of 40 topics, with hyperparameter α = .01.
7. Examined distribution lists of top words (words most unique to each topic) and top documents (scenes identified by the model as most representative of each topic or theme) to qualitatively analyze the content and thematic meaning of individual topics and themes.
   1. Each topic was coded and labelled by two independent coders examining the word lists and the full scenes to identify common content that defined each.

The final model is comprised of 188 films broken into 11,900 scenes. It has a vocabulary of 29,324 unique words (trimmed to 12,241 words after sparsity reduction; see Maier et al., 2018).

**References**

Blei, D. M., Ng, A. T., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

BoxOfficeMojo (n.d.) *Top grossing romantic comedy movies at the box office*. https://www.boxofficemojo.com/genres/chart/?view=main&sort=gross&order=DESC&pagenum=5&id=romanticcomedy.htm

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, *70*(6), 066111. DOI: 10.1103/PhysRevE.70.066111.

Jockers, M. L. (2014). *Text Analysis with R for students of literature*. Springer.

Maier, D., Waldherr, A., Miltner, P., Wiedeman, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods & Measures*, *12*, 93-118. DOI: 10.1080/19312458.2018.1430754.

Murtagh, F., Ganz, A., & McKie, S. (2009). The structure of narrative: the case of film scripts. *Pattern Recognition*, *42*(2), 302-312.

Walter, D., & Ophir, Y. (2019). News Frame Analysis: An Inductive Mixed-method Computational Approach. *Communication Measures & Methods*, *13*(4), 248-266. DOI: 10.1080/19312458.2019.1639145.

Walter, D., Ophir, Y., & Jamieson, K. H. (2020). Russian Twitter Accounts and the Partisan Polarization of Vaccine Discourse, 2015–2017. *American Journal of Public Health*. Epub ahead of print 15 April 2020. DOI: 10.2105/AJPH.2019.305564.