Received: 24 July 2020 | Revised: 8 April 2021 | Accepted: 11 May 2021

DOI: 10.1002/asi.24525

RESEARCH ARTICLE

JASIST WILEY

# Genre analysis of movies using a topic model of plot summaries

**Paul Matthews** | **Kathrina Glitre**

Department of Computer Science and Creative Technologies, UWE Bristol, Bristol, UK

**Correspondence**
Paul Matthews, Department of Computer Science and Creative Technologies, UWE Bristol, Coldharbour Lane, Bristol BS16 1QY, UK.
Email: paul2.matthews@uwe.ac.uk

**Abstract**

Genre plays an important role in the description, navigation, and discovery of movies, but it is rarely studied at large scale using quantitative methods. This allows an analysis of how genre labels are applied, how genres are composed and how these ingredients change, and how genres compare. We apply unsupervised topic modeling to a large collection of textual movie summaries and then use the model's topic proportions to investigate key questions in genre, including recognizability, mapping, canonicity, and change over time. We find that many genres can be quite easily predicted by their lexical signatures and this defines their position on the genre landscape. We find significant genre composition changes between periods for westerns, science fiction and road movies, reflecting changes in production and consumption values. We show that in terms of canonicity, canonical examples are often at the high end of the topic distribution profile for the genre rather than central as might be predicted by categorization theory.

## 1 | INTRODUCTION

Theories and approaches to genre are often based on limited, handpicked examples. With the advent of digital methods, there is the opportunity for larger scale mapping and analysis of cultural works. While genre is often used as a categorical label in machine learning model building, few studies looking computationally at cultural data have used the available methods to study the structure and dynamics of genre itself.

A key challenge is to understand the human construct of genre and the social labeling of a corpus through quantitative methods. This coupling of machine and folk approaches may be particularly powerful in illustrating which aspects of culture are transparent in a certain form of representation and which are less tangible or conceptual. It also provides coarse level corroboration or refutation for both cultural and cognitive theory that has hypothesized how genre categories are related and exemplified. This study takes the above approach for illustrating patterns and practice in genre classification of movies using a labeled corpus of plot summaries and by applying topic modeling.

From our recovered topic distributions, we show how distinctive genre labels are and how genres may be mapped using their topic proportions. Next, we compare temporal shifts in the topic makeup of genre, showing some statistically significant trends in lexical ingredients over time within westerns, science fiction and road movies. Finally, we use topical distributions to examine the phenomenon of genre canon. As category exemplars, we might expect these to behave as prototypical or

median examples within the topic distributions. However, we show that canonical examples tend to be positioned at the more extreme end for genre signatures.

We proceed with an introduction to conceptions of genre as related to film and to how these might be examined structurally. Then, we introduce theories of genre change/evolution. We go on to relate the idea of canonicity in genre to prototype theory in classification. Finally, we consider the previous application of machine learning to genre in cultural collections.

## 1.1 | The nature of genre

Genre is something of a common cultural consensus: "genre is what we collectively believe it to be" (Tudor, 1976, p. 122). Genre can be considered an open set, with the potential to morph via new membership at any time (Dimock, 2007). The fact that people find it so hard to agree on whether particular examples are instances of a genre or not may indeed indicate that it is less about presence or absence of specific properties and more about socially constructed traditions, with an upstream (production) and downstream (consumption) stakeholder component (Evnine, 2015). With the continuous creation of new genres and the replication of generic tropes, genre typifies Cawelti's well-known aphorism that art can be understood on a continuum between invention and convention (Grant, 2007).

Genre certainly has different functions for different user groups, with the needs of industry and theoretical/scholarly authors being very different (Neale, 2007). The pragmatic, communicative aspects of genre are frequently underemphasized (Furstenau, 1995). While critical analysis may tend toward singling out genre and viewing it as monohierarchical, publicists are often keen to emphasize multiple genre appeal and therefore polyhierarchy (Geraghty & Jancovich, 2008). Certainly, there is some evidence that a film's genre is a motivating factor in cinema visits and is still found—albeit in more granular or hybrid forms—in modern streaming services.

## 1.2 | The structural study of genre

A key question for the present kind of study is whether genre can be studied using a large dataset and relatively surface level features. Certainly, quantitative methods need to prove their value in adding to knowledge rather than repeating well-known findings (Henrichs, 2019; Schmidt, 2016). However, as long ago as 1976, Tudor, while noting that structuralist approaches are "rarely satisfactory," went on to say that the methods had borne some fruit and enabled analysts to "break down the regal isolation of the film as basic unit of analysis" and identify previously unnoticed patterns (Tudor, 1976, p. 125). A recent review concludes that with formal methods (for musical tastes) "there is coherence and consistency that cannot be grasped by merely identifying and classifying individual tastes" and that various levels of interdependence, relations and social patterns can be revealed (Edelmann & Mohr, 2018, p. 2). On a similar note, abstraction and deformation of the original text may open the way to new and innovative analyses (Henrichs, 2019).

In literary studies, computational techniques have been applied as a telescope to "distant reading," in contrast to the microscope of close reading, losing the detail and experience of engagement through reading (Eve, 2019). Distance may however confer the advantages of reducing distractions and placing the emphasis on interconnections and relationships (Moretti, 2005).

A natural objection for cultural works is that structural and content-based analysis may only target "outer form" (Grant, 2007) and neglect important aspects of genre conventions such as mise-en-scène, style, iconography and specific dialogue. While acknowledging this shortcoming, we might also note the risks of generalization about genre from a few handpicked examples. Such individual theorizing about genre is necessarily selective in the examples it uses (Neale, 2007).

Genres typically do have conventions or motifs, such as gunfights and villains in westerns, that should be recognizable from explicit descriptions (Tudor, 1976). These features may become audience expectations which filmmakers may then exploit in order to "break the rules" for the genre (Tudor, 1976). So while noting that many aspects of a rounded genre experience will not be recoverable with computational (and textual, plot-based) methods, we might expect that certain genres will have a recognizable set of features.

## 1.3 | Genre lifecycles

A received lifecycle model in the early film genre literature is that a genre progresses through primitive, classical, revisionist, and parodic periods. At first, according to this model, the themes are novel, but conventions are gradually established (primitive), after which there is a period in which the values and expectations are shared between audience and makers (classical). Thereafter, conventions start to be challenged and subverted (revisionist) after which the previously celebrated conventions are exaggerated or ridiculed (parodic).

Neale (2007) presents some additional parallel/competing models: that of growth, flowering and decay; progression toward self-conscious formalism; or the Russian formalist approach (evolutionary-inspired) of seeing genre as a competition between alternatives or a continual process of "canonization, automation, and reshuffling" (Neale, 2007). This final view seems to accommodate a more dynamic and parallelized process of rise and fall. The linear evolution view may be too simplistic and instead there may be multiple developmental cycles (Geraghty & Jancovich, 2008).

Genres can be seen as historical constructs, where works are understood in the context of their time. (Underwood, 2016a). As such, early theorists speculating about "classical periods" within a genre may have been lacking a long view (Grant, 2007). In deeming "Great Train Robbery," the first classic Western, for example, theorists may not be aware that at the time audiences were less familiar with the term and its popularity may have been more due to the way it combined crime and railway themes. As Neale (2007, p. 169) puts it, "Genres are inherently temporal: hence, their inherent mutability on the one hand and their inherent historicity on the other."

So we should be aware that generic terms may be imposed, retrospectively (Geraghty & Jancovich, 2008). However, despite this concern, there has been a long enough period of well-documented movie production and consumption for there to be a good amount of evidence to detect temporal shifts in how genre labels are construed and applied.

## 1.4 | Prototypes and canons

Rosch's prototype theory provided the robust experimental finding that people tend to quote similar examples of goodness-of-fit for a category, with these best examples referred to as prototypes (Lakoff, 1987). Against classical category theory, prototype theory thus "defends a non-criterial conception of categorial structure" (Geeraerts, 1989, p. 588). That is, the boundaries of categories are fuzzy and they rely on flexible and contingent associations. Film genre theorists have termed genres "family resemblances" in a close echo of the prototype theory (Underwood, 2016b). While some academic definitions of genre may tend to see them as self-contained and mutually exclusive, a more flexible category-based view would see them as being mixed or combined with ease (Geraghty & Jancovich, 2008).

Key works within a genre are often used to represent key features or tendencies (Geraghty & Jancovich, 2008). Where these features are explicit and recoverable via information extraction, we might be able to use canonical examples to understand whether these show prototypical characteristics.

## 1.5 | Genre, metadata, and discovery

A key practical application of genre is in organizing a body of work and in finding similar examples. The "semantic gap" between machine-recoverable features of multimedia and human retrieval expectations means that human-applied metadata remains important for this purpose, though the two may work productively together for discovery (MacFarlane, 2016).

Genres and subjects (what films "are" and what they are "about") may be conflated in the cataloguing processes (de Esteban, 2012). As a result, at the British Film Institute, cataloguers are advised to select just one or two genres but potentially more subjects. The example of Dr Zhivago is used to illustrate a work that should be assigned to the romance and drama *genres* but which includes war, action and adventure *subjects* (Fairbairn, pers. comm.).

Netflix created "altgenres" from composite human-applied tags that describe the region, feeling, genre, setting, subject and period among other things (Lawrence, 2015). These are just a part of an interface that is also driven by user interaction behavior and content availability (Madrigal, 2014). Lawrence (2015) notes the tension between the platform's need to drive engagement and the viewers' own agency, and wonders if these dimensions can really cater to the user's own aesthetic preferences.

Certainly, digital discovery methods are well suited to more detailed description and fine-grained characterization. In using methods that achieve a kind of digital genotype from text content, it should be feasible to develop new supplementary navigation and discovery opportunities.

## 1.6 | Machine learning and genre

With movie data, the most commonly addressed tasks have been the prediction of genre from content-based features and the use of genre in the development of user recommendation algorithms. Ertugrul and Karagoz (2018) for instance, used deep learning to model short plot summaries labeled by genre, with 1,600 examples of each genre. After training, they obtained F-scores for genre prediction—with comparable precision and recall—of between 61 and 77% depending on the genre predicted (horror scoring highest, thriller lowest). This is quite

impressive, as the plot summaries used were quite short. Chao and Sirmorya (2016) used topic modeling on movie scripts, generating 190 topics from a corpus of 1,094 scripts. They then used cosine similarity to predict genre based on nearness to labeled examples, gaining an F1 accuracy of 0.49, with much higher recall than precision. That is, the prediction seemed to find the majority of the scripts in the genre being predicted, but along with those from a lot of other genres (false positives).

Few studies looking computationally at movie data have used the methods to actually study genre itself. For literary texts, Underwood (2016b) used logistic regression to predict genre based on the top 10,000 words across a set of texts, achieving 70–93% prediction accuracy, depending on genre. Those with very distinctive keywords, such as detective and science fiction, achieved the highest accuracy. Underwood was able to show, to some extent, that genre was relatively consistent across the time window he was studying (19th and 20th centuries), a challenge to theorists who proposed a waxing and waning within shorter lifespans.

These studies are something of a justification for the identification of genre from language and narrative explication. The subtleties of both gross misclassifications and the positioning of individual cases are equally illuminating from a genre representation and discovery perspective. Much of the above-quoted work uses supervised techniques, relying on training data or ground truth, with pre-labeled examples or a matrix of genres, user views and existing ratings. It is also aimed at the prediction of labels or rating, rather than being used as a tool to visualize and understand patterns within the data itself. In contrast, the present study is based around the use of an unsupervised approach, where the corpus is differentiated through probabilistic pattern recognition. These *machine-discernible* patterns can then be used to compare and contrast with the *human-generated* metadata in order to understand how genre relates to structural and temporal features of the corpus.

## 2 | METHODS

### 2.1 | Dataset

We used the CMU movie summary corpus (Bamman, O'Connor, & Smith, 2014) of 42,306 summaries from Wikipedia with aligned metadata from Wikidata (formerly Freebase). This was originally used in a study to understand character types from lexical features (Bamman, O'Connor, & Smith, 2013). The origin of the summaries and genre classifications are rather obscure, but are assumed to be a mixture of authoring by Wikipedia/Wikidata editors and preexisting lists and summaries that have been imported. Many of the summaries, for instance, are common to a range of online movie databases, so it is unclear where they first originated, and we can expect a good proportion to be based on the original promotional material from the studio itself. The genre classification will be similarly mixed in origin and may frequently be due to the judgment of a single editor, as we know that the first entries on user-generated content repositories often tend to persist (e.g., see Gazan, 2015).

It should be noted that the representation of "genre" in this dataset is somewhat looser than what would be considered useful within film studies. While it does include some well-established academic labels such as "Western" and "Film Noir," it also has a number of terms that are either outmoded, xenophobic or so generic as to be not considered useful for theorizing (e.g., "Black-and-White" or "Japanese Films").

As the present study required a reasonable length of summary and we were also interested in the release year of movies, we filtered the CMU dataset to remove records without a year and with a summary length of less than 400 characters (Wikipedia guidelines state that summaries should be 400–800 characters in length). This filtering step resulted in a dataset of 32,758 movies.

#### 2.1.1 | Characteristics

The mean summary length of the resultant data was 2,143 characters, with a median of 1,487 (hence, a skew toward shorter summaries and a smaller proportion of very long examples). In the metadata, there were a total of 358 genre labels applied, with an average of 3.88 genres per film, ranging from 1 to 17. Table 1 gives the numbers for the top 20 most widely applied genres. At the lower end were a number of very niche genres (e.g., "Clay animation," "Werewolf fiction," "Cyberpunk"). These typically were only applied to a very small number of works and as such were not extensively referenced in the analysis.

In understanding the corpus, it is also instructive to note the primary language of the work. While the summaries were taken from English Wikipedia, the original work itself was often not in English (Table 1).

### 2.2 | Topic modeling

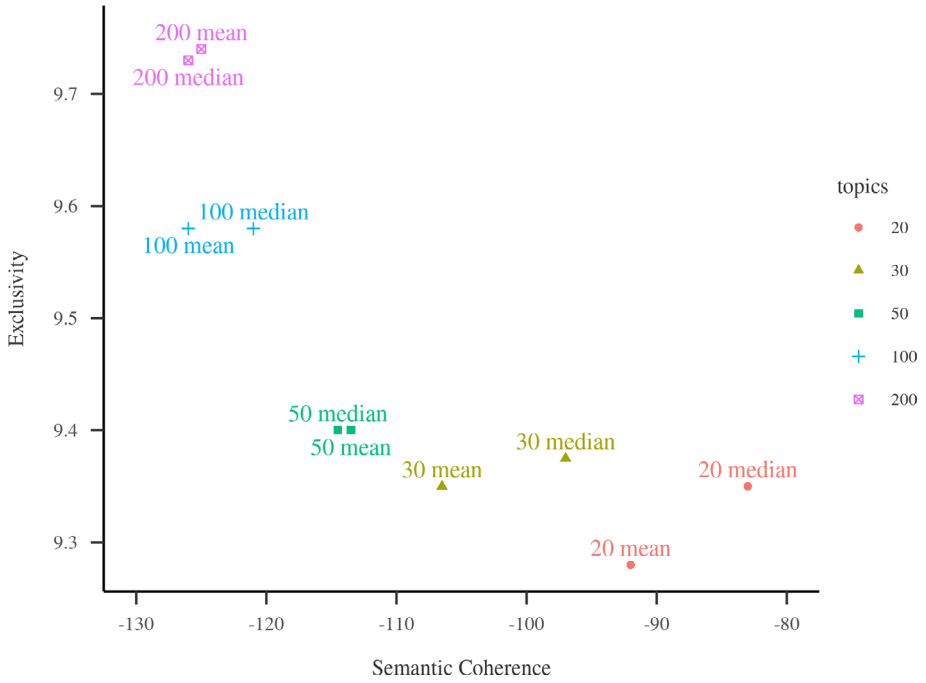
#### 2.2.1 | Overview of the method

Topic modeling is a probabilistic approach that works backward from the assumption that a corpus of documents can be modeled using a discrete set of topics

**TABLE 1** Top 10 most prevalent genres and languages in the dataset (note that films can belong to more than one genre)

| Genre | Count (genre) | Language | Count (language) |
|---|---|---|---|
| Drama | 15,389 | English | 21,978 |
| Comedy | 8,705 | Hindi | 1,438 |
| Romance film | 5,642 | Spanish | 1,140 |
| Thriller | 5,494 | Tamil | 877 |
| Action | 4,893 | Japanese | 706 |
| World cinema | 4,623 | French | 653 |
| Crime fiction | 3,482 | Malayalam | 588 |
| Horror | 3,326 | German | 528 |
| Action/adventure | 3,189 | Telugu | 502 |
| Indie | 2,985 | Cantonese | 440 |

**FIGURE 1** Topic diagnostics




(essentially made up of words likely to appear in that topic), where each document instance contains some mixture of these overall topics. Topic-word and document-topic distributions are estimated using iterative sampling or variational inference algorithms. These procedures use the (user-specified) number of required topics and the token distribution across the corpus to infer the distributions, satisfying constraints over the exclusivity of topics and the fine-grainedness of the topics within the documents. For more details of the modeling and estimation logic, see Blei, 2012a, 2012b; Blei, Ng, & Jordan, 2003).

Since the introduction of the approach, a number of variants of the original formulation have been developed, many providing "semisupervised" flavors, where additional input variables are used to establish prior expectations as to topic content and/or distribution—for example: disciplines in analyzing scientific papers (D. M. Blei & Lafferty, 2007); authors to guide the modeling of their documents (Giaquinto & Banerjee, 2018). The present study (to a large extent) deliberately did *not* use the semisupervised approach across all genres in order to (a) not to impose a prior expectation that topic distributions would be correlated with genre; and (b) provide a unified set of topics that could be used to shed light on the composition and structure of genres.

### 2.2.2 | Current approach

We used the Structured Topic Model (STM) package in R (Roberts, Stewart, Tingley, & Airoldi, 2013) to develop

our topic models. This package provides the full pipeline toolset, including corpus preparation, model estimation and analysis tools once the model has been constructed. Importantly, it also uses a spectral initialization procedure—the starting point for estimating the model—based on a matrix factorization of the word co-occurrence matrix, which provides much better model reliability, lack of reproducibility having been a criticism of topic modeling (Agrawal, Fu, & Menzies, 2018).

To prepare the corpus we applied the standard stop word filtering and removed words with 10 occurrences or fewer. Given that many movie summaries also include character names which were not informative for our analysis, we filtered out the most common names using a (US-centric) baby names corpus (Wickham, 2019) and a dataset of Indian names (Patil, 2019).

In topic modeling, a key decision is the number of topics to model and here model validity, model characteristics and/or user judgment can guide the choice. We used STM's ability to compare the characteristics of topic exclusivity—the harmonic mean of words jointly ranked by frequency and exclusivity—and semantic coherence—how likely the top words are to appear together in a document, a proxy for understandability (Mimno, Wallach, Talley, Leenders, & McCallum, 2011). To ascertain these parameters, we estimated models at 20, 30, 50, 100, and 200 topics (the *K* value) and Figure 1 illustrates the outcome for coherence and exclusivity. Based on this

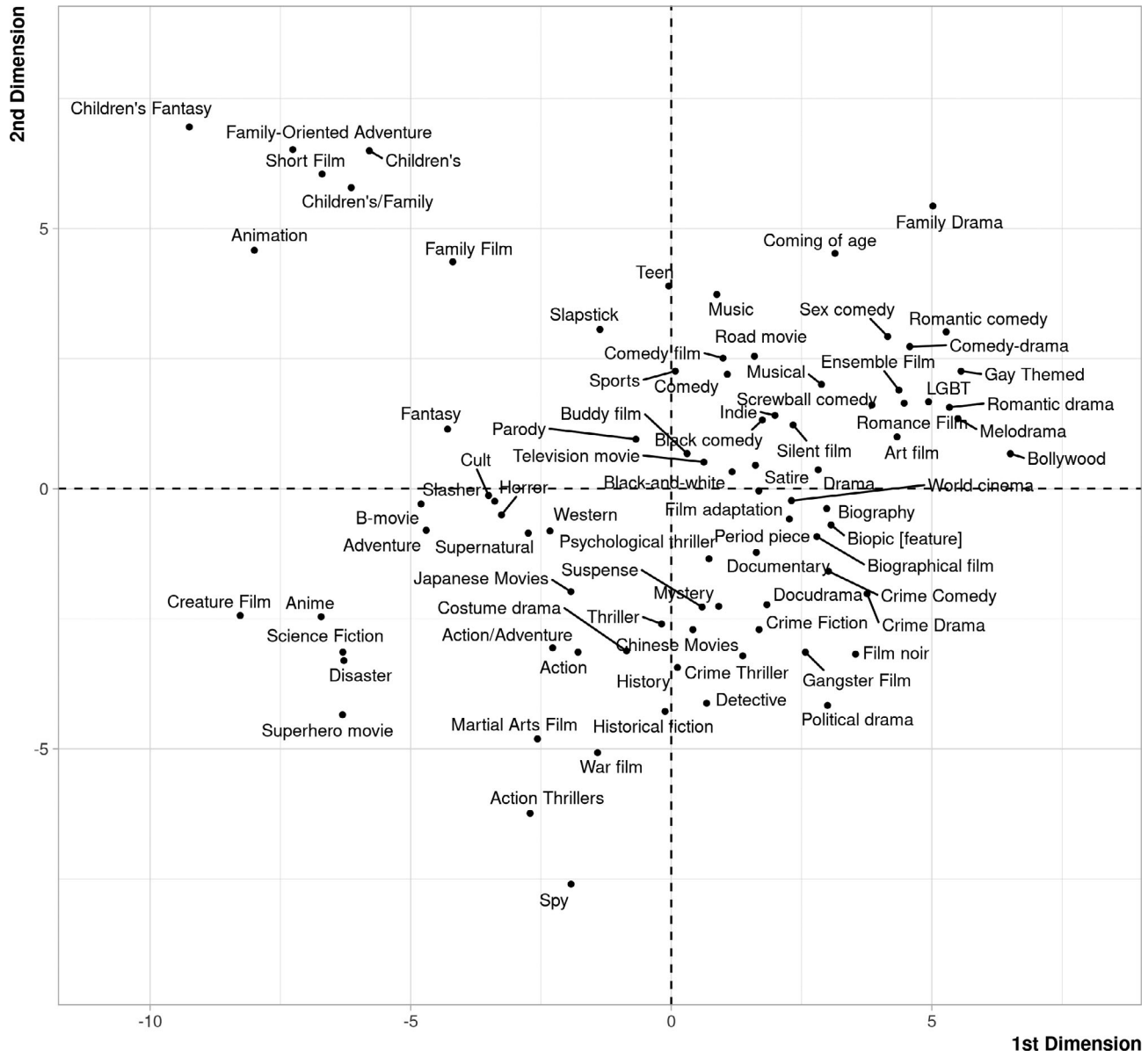



FIGURE 2 Principal Components 1 and 2

investigation, a $K$ value of 70 was used as representing the best tradeoff between exclusivity and coherence.

Having estimated the vanilla $K = 70$ model, we could combine the matrix theta of estimated topic proportions per document with the movie metadata and use this as the basis for further analysis.

### 2.3 | Selection of canonical/classic examples

In order to test the distribution of examples viewed as genre classics, we selected instances for each genre using a search engine selection method. Five examples were selected from the top 30 search engine results using the phrase "classic (...genre...) movies." E-commerce sites were discounted. The first page for each distinct publisher was used. Many pages rank the top 100 or 20 examples in the genre and from these the top five were chosen as candidates. The five examples used were the candidates mentioned the most frequently across eligible sites (usually seven to nine distinct publishers). Ties were selected from using a random draw. These sites tended to include both editorial selections and crowdsourced rankings (e.g., the *Rotten Tomatoes* review aggregator).

## 3 | RESULTS

### 3.1 | Genre distinctiveness and lexical motifs

#### 3.1.1 | Dimension reduction

As our result set varied over 70 dimensions (one for each topic), we first sought to reduce dimensions to see which topics most contributed to the overall variation within the corpus. To understand this distinctiveness of genres, we therefore computed the principal components on the median topic distributions of genres with more than 200 instances.

The first three dimensions explained 46% of the variance in topic composition and 10 dimensions were needed to explain 75%. Figure 2 plots the genres on the first two dimensions. Here, the genres farthest from the origin are most distinctive in the first dimension, and those close together have a similar topic composition. There is a distinctive cluster for children's films in the top left quadrant, with the remainder more densely grouped. We can see some genres with commonalities to children's films, though for a more adult audience in the bottom left (e.g., superhero movies, anime). The lower right-hand side are certainly more adult themes and crime-focused. At the top right are potentially more "harmless" genres, with some teen and family genres.

Tables 2–4 show top words for the most significant topics in the first three dimensions, and hence those topics that most easily distinguish between genres

**TABLE 2** Significant topics in dimension 1

| Top terms in topic | Correlation | *p*-Value |
|---|---|---|
| madam, spanish, brothel, Spain, del | 0.82 | .00 |
| woman, man, old, meet, name | 0.82 | .00 |
| new, work, job, compani, busi | 0.78 | .00 |
| life, live, time, one, feel | 0.74 | .00 |
| water, boat, sea, fish, swim | −0.72 | .00 |
| bear, camp, mountain, hunt, wood | −0.71 | .00 |
| island, group, kill, find, attack | −0.71 | .00 |
| dog, cat, bird, back, get | −0.68 | .00 |
| husband, wife, affair, suicid, attempt | 0.67 | .00 |
| play, danc, perform, show, act | 0.67 | .00 |

**TABLE 3** Significant topics in dimension 2

| Top terms in topic | Correlation | *p*-Value |
|---|---|---|
| christma, eve, santa, toy, doll | 0.93 | .00 |
| assassin, kill, infect, virus, puppet | −0.82 | .00 |
| boy, man, anim, circus, lion | 0.76 | .00 |
| kid, tree, get, friend, flower | 0.72 | .00 |
| shadow, bean, gwen, deed, sheep | 0.71 | .00 |
| valley, find, ball, robinson, bloom | 0.69 | .00 |
| prison, captain, ship, ladi, escap | −0.63 | .00 |
| tell, get, say, ask, leav | 0.59 | .00 |
| farm, egg, barn, farmer, red | 0.58 | .00 |
| daffi, get, duck, porki, stoog | 0.58 | .00 |

**TABLE 4** Significant topics in dimension 3

| Top terms in topic | Correlation | *p*-Value |
|---|---|---|
| mrs, hotel, money, pay, miss | 0.79 | .00 |
| polic, offic, arrest, escap, crimin | 0.75 | .00 |
| bank, robberi, plan, money, robber | 0.72 | .00 |
| store, chicago, bar, drink, club | 0.60 | .00 |
| help, tri, take, two, fall | 0.57 | .00 |
| murder, case, investig, crime, suspect | 0.52 | .00 |
| car, drive, truck, train, driver | 0.52 | .00 |
| count, baron, howev, fortun, return | 0.51 | .00 |
| gang, drug, angel, street, gangster | 0.50 | .00 |
| villag, child, wife, pregnant, husband | −0.49 | .00 |

TABLE 5 Accuracy of genre prediction based on topical proportions

| Genre | TM70 P | TM70 R | TM70 F1 | TFIDF F1 | *N* |
|---|---|---|---|---|---|
| Martial arts film | 0.96 | 0.94 | 0.95 | 0.60 | 631 |
| Western | 0.95 | 0.89 | 0.92 | 0.72 | 786 |
| War film | 0.87 | 0.94 | 0.90 | 0.74 | 1,315 |
| Animation | 0.83 | 0.96 | 0.89 | 0.63 | 1,887 |
| Documentary | 0.91 | 0.86 | 0.88 | 0.60 | 863 |
| Sports | 0.94 | 0.82 | 0.88 | 0.56 | 581 |
| Slasher | 0.98 | 0.78 | 0.87 | 0.13 | 554 |
| Horror | 0.75 | 0.98 | 0.85 | 0.92 | 3,326 |
| Science fiction | 0.75 | 0.98 | 0.85 | 0.76 | 2,034 |
| Fantasy | 0.77 | 0.95 | 0.85 | 0.53 | 1,770 |
| Chinese movies | 0.89 | 0.81 | 0.85 | 0.54 | 826 |
| Supernatural | 0.91 | 0.80 | 0.85 | NA | 518 |
| Children's/family | 0.87 | 0.81 | 0.84 | NA | 690 |
| Crime thriller | 0.76 | 0.91 | 0.83 | 0.46 | 1,420 |
| Psychological thriller | 0.76 | 0.90 | 0.82 | NA | 984 |
| Cult | 0.85 | 0.80 | 0.82 | 0.08 | 645 |
| Action/adventure | 0.67 | 0.98 | 0.80 | 0.70 | 3,189 |
| Romantic comedy | 0.70 | 0.93 | 0.80 | 0.46 | 1,831 |
| Coming of age | 0.86 | 0.75 | 0.80 | NA | 669 |
| Suspense | 0.91 | 0.71 | 0.80 | NA | 579 |
| Teen | 0.82 | 0.77 | 0.79 | 0.13 | 763 |
| Romantic drama | 0.65 | 0.99 | 0.78 | 0.55 | 2,316 |
| Mystery | 0.69 | 0.91 | 0.78 | 0.50 | 1,748 |
| Adventure | 0.63 | 0.98 | 0.77 | 0.73 | 2,776 |
| Short film | 0.69 | 0.88 | 0.77 | 0.73 | 2,047 |
| Action | 0.61 | 0.99 | 0.75 | 0.78 | 4,893 |
| Family film | 0.63 | 0.92 | 0.75 | 0.73 | 2,654 |
| Period piece | 0.65 | 0.82 | 0.73 | NA | 1,197 |
| Comedy-drama | 0.67 | 0.80 | 0.73 | NA | 1,065 |
| Family Drama | 0.82 | 0.65 | 0.73 | NA | 671 |
| Indie | 0.56 | 0.99 | 0.72 | 0.56 | 2,985 |
| Black-and-white | 0.56 | 1.00 | 0.72 | 0.49 | 2,926 |
| LGBT | 0.77 | 0.68 | 0.72 | 0.41 | 711 |
| Crime fiction | 0.56 | 0.97 | 0.71 | 0.76 | 3,482 |
| Musical | 0.58 | 0.93 | 0.71 | 0.52 | 2,066 |
| Film adaptation | 0.61 | 0.85 | 0.71 | NA | 1,128 |
| Thriller | 0.54 | 1.00 | 0.70 | 0.74 | 5,494 |
| Japanese movies | 0.69 | 0.72 | 0.70 | 0.47 | 1,034 |
| Romance film | 0.53 | 0.99 | 0.69 | 0.71 | 5,642 |
| Drama | 0.51 | 1.00 | 0.68 | 0.67 | 15,389 |
| World cinema | 0.52 | 1.00 | 0.68 | 0.67 | 4,623 |
| Black comedy | 0.68 | 0.68 | 0.68 | NA | 751 |
| Comedy | 0.51 | 0.99 | 0.67 | 0.68 | 8,705 |

TABLE 5 (Continued)

| Genre | TM70 P | TM70 R | TM70 F1 | TFIDF F1 | *N* |
|---|---|---|---|---|---|
| Silent film | 0.74 | 0.62 | 0.67 | 0.02 | 723 |
| Comedy film | 0.56 | 0.79 | 0.66 | 0.02 | 1,140 |
| Parody | 0.69 | 0.59 | 0.64 | NA | 696 |
| Television movie | 0.78 | 0.39 | 0.52 | NA | 557 |
| Bollywood | 0.83 | 0.15 | 0.25 | 0.33 | 1,013 |
| Biography | 0.83 | 0.05 | 0.09 | NA | 524 |
| Biographical film | NA | 0.00 | NA | NA | 508 |

*Note:* P = precision; R = recall.

TABLE 6 Science fiction films by period

| Period | Years | Count |
|---|---|---|
| 2 | 1925–1945 | 53.00 |
| 3 | 1947–1968 | 285.00 |
| 4 | 1969–1990 | 541.00 |
| 5 | 1991–2013 | 1,143.00 |

explaining much of the variation between points in the plot. We see that sex/crime, adult, and professional themes having a positive correlation with the first dimension, whereas outdoor and animal themes correlate negatively (i.e., these topics are significantly lower in those genres at the low end of the dimension). In the second dimension, Christmas topical content is a highly significant determinant of positive genre location along with other childhood-related content. Spy and prison topics are negatively correlated in this dimension. In the third dimension, gangster and crime themes dominate in the positive direction, whereas family contributes negatively.

### 3.1.2 | Genre prediction power

The relative distinctiveness of signature topics was then further investigated to understand their performance in prediction. A simple logistic model was trained for each of the top 82 genres (those with at least 200 instances) using the full dataset, with the exception of a held out test of 100 positive and 100 negative examples, used to assess performance. Results are shown in Table 5, below. We see the most predictable genres toward the top, which correspond with the motif-heavy genres. Those further down include genre labels relating to the format more than the content (e.g., silent film, television movie) or very broad categories (e.g., drama/world cinema).

To understand the relative usefulness of topics compared to token-level features, we compared the prediction accuracy with a TF/IDF feature set. We find that, with a few exceptions, the TF/IDF prediction does not give the same performance as the 70 topic prediction. This may be because the topics balance term prevalence with term exclusivity, while TF/IDF focuses just on exclusivity, resulting in high precision but low recall. Table 5 also shows that there is no discernable relationship between the number of examples in each category and the genre predictability. Hence, this seems to have more to do with the vocabulary makeup of the genre plots.

## 3.2 | Temporal effects on topic composition

Using the movie release year from the metadata, topic composition could be compared between periods within a genre. This was done for science fiction, westerns and road movies and results are shown below. Tables 6, 8, and 10 show the periods used and the count of examples within each. Figures 3–5 show the mean topic proportions over the periods (for topics with at least 2.5% contribution across periods). Finally, Tables 7, 9, and 11 give the results of analysis of variance tests for the (log-transformed) topic proportions.

For science fiction, we see growth and then shrinkage of the signature "earth, ship, crew, space, destroy" topic, but gradual growth of a second similar signature "alien, human, robot, destroy, earth." This may indicate a shift from outward looking space exploration- to inward, earth invasion-related story lines. Two example excerpts are shown below. We also see growth of the fantasy topic "battle, kill, dragon, sword, attack," indicating the close relation between the science fiction and fantasy genres, with elements of both more common in more recent works. This was the most marked topical shift across periods.

TABLE 7 ANOVA of science fiction topic trends

| | *df* | Sum Sq | Mean Sq | *F* value | Pr (>F) |
|---|---|---|---|---|---|
| earth, ship, crew, space, destroy | 1 | 30.30 | 30.30 | 48.23 | .00 |
| island, group, kill, find, attack | 1 | 18.22 | 18.22 | 28.99 | .00 |
| tell, get, say, ask, leav | 1 | 17.02 | 17.02 | 27.08 | .00 |
| battl, kill, dragon, sword, attack | 1 | 15.09 | 15.09 | 24.02 | .00 |
| alien, human, robot, destroy, earth | 1 | 10.33 | 10.33 | 16.44 | .00 |
| life, live, time, one, feel | 1 | 2.15 | 2.15 | 3.42 | 0.06 |
| kill, find, shoot, bodi, attack | 1 | 1.31 | 1.31 | 2.09 | 0.15 |
| agent, bomb, plane, kill, team | 1 | 1.16 | 1.16 | 1.85 | 0.17 |
| world, use, power, machin, time | 1 | 0.51 | 0.51 | 0.81 | 0.37 |
| Residuals | 2,012 | 1,264.23 | 0.63 | NA | NA |

Abbreviation: ANOVA, analysis of variance.

TABLE 8 Westerns by period

| Period | Years | Count |
|---|---|---|
| 2 | 1926–1947 | 133.00 |
| 3 | 1948–1969 | 332.00 |
| 4 | 1970–1991 | 156.00 |
| 5 | 1992–2013 | 141.00 |

"The year is 1980, and the film opens with the launch of the JX-1 Hayabusa spaceship into outer space. The ship, originally sent to collect data on Saturn, has its course diverted to investigate the mysterious star Gorath."—Signature Topic 1 (exploration), Gorath (1962).

"Teenage brothers Nick and Tyler come across a UFO that crashed near their town. Soon after evading the government, mainly agent Armstrong, and keeping the UFO in seclusion, Tyler goes through dramatic physical changes and gains superhuman abilities."—Signature Topic 2 (invasion), Skyrunners (2009).

In the case of westerns, while the signature topic "town, men, hors, sheriff, ride" is a consistently high topical feature, there is a significant decrease across periods, this decrease being echoed by reductions in a "war" topic and a "business" topic "new, work, job, compani, busi." The latter finding—a decrease in work themes—seems to be linked to storylines featuring oil, the railway and also enterprise or job reasons being the reason for lead characters to travel west:

"Jonathon Tibbs, son of a family of English gunsmiths, has no interest in the business and prefers inventing gadgets, in particular a steam-powered horseless carriage... He reads in his newspaper about the wide use of guns in the American West of the 1880s, and decides to go there himself to sell firearms to the locals."—Signature Topic 3 (work), The Sheriff of Fractured Jaw (1958).

"Drifter cowboy Jim Garry receives a job offer by mail from smooth-talking Tate Riling. Garry rides into an Indian reservation and finds himself in the middle of a feud between cattle ranchers and homesteaders."—Signature Topic 3 (work), Blood on the Moon (1948).

For road movies, the signature "car, drive, truck, train, driver" topic did not vary significantly over the periods. In contrast, there was a decrease in "bank, robberi, plan, money, robber" from its peak in the late 60s and 70s and an accompanying increase in dialogue content ("tell, get, say, ask, leav") and the "self-actualization" topic ("life, live, time, one, feel") into the 21st century. These changes clearly reflect a change from road movies that portray malfeasance to those that use the journey theme to portray personal growth and renewal.

"The trip turns out to ultimately cure Victor's brooding disposition toward life and shows him why his father became an alcoholic, was abusive, and eventually left their family."—Signature Topic 21 (self-actualization), Smoke Signals (1996).

"But what starts out as the tour from hell turns into a meaningful journey, with an unexpected series of revelations that will change all of their lives"—Signature Topic 21 (self-actualization), Everything Is Illuminated (2005).

## 3.3 | Location of canonical and revisionist examples

Genres often have a number of films considered canonical, or embodying the themes of the genre. To investigate the topic proportions of such examples, we took their topic proportion and overlaid them on the distribution of the topics as a whole. Examples were drawn from the

**TABLE 9** ANOVA of western topic trends

| | *df* | Sum Sq | Mean Sq | *F* value | Pr (>F) |
|---|---|---|---|---|---|
| town, men, hors, sheriff, ride | 1 | 16.33 | 16.33 | 17.77 | .00 |
| new, work, job, compani, busi | 1 | 13.25 | 13.25 | 14.42 | .00 |
| war, soldier, armi, german, command | 1 | 4.03 | 4.03 | 4.38 | 0.04 |
| kill, find, shoot, bodi, attack | 1 | 1.42 | 1.42 | 1.55 | 0.21 |
| tell, get, say, ask, leav | 1 | 1.15 | 1.15 | 1.25 | 0.26 |
| Residuals | 756 | 694.75 | 0.92 | NA | NA |

Abbreviation: ANOVA, analysis of variance.

**TABLE 10** Road movies by period

| Period | Years | Count |
|---|---|---|
| 3 | 1966–1979 | 41.00 |
| 4 | 1981–1996 | 56.00 |
| 5 | 1997–2012 | 126.00 |

Wikipedia genre pages (Contributors, 2020a, 2020b; Contributors, 2020c).

For westerns, the examples are shown in Figure 6. We see that for two oft-quoted examples, "High Noon" and "Stagecoach," topic proportions for the signature topic are in the upper quartile and upper whisker respectively in terms of distribution. In contrast, proportions of that topic for "A Man Called Horse"—from the revisionist period— is at the top of the lower whisker (quartile 1–1.5 × the interquartile range). For "Brokeback Mountain," a movie where there is considerable contention as to its genre, and where the director did not have a clear intention to make a western (Spohrer, 2009), the signature elements are also in the lower whisker.

Road movie examples are somewhat more extreme than westerns in signature content (Figure 7). "Easy Rider" and "Two-Lane Blacktop" are outliers for road and vehicle content. "Easy Rider" was perhaps the first work to be classed by reviewers as a road movie (Hurault-Paupe, 2015). "Thelma and Louise," a more recent work is found in the upper whisker. "Little Miss Sunshine," a recent example with more content around the character relationships, is close to the median for signature content.

In the systematic analysis of canonical movies against signature topics, as selected by the search engine ranking and tabulation procedure, we see that classic examples are frequently, though not always, ranked very highly (in the top 10%) for their signature topics. In cases where there are more than one signature topic within the genre, the examples tend to be leaders for at least one—for example, Rocky in the sports genre is an outlier against the fight-related topic but not against the team sports-related topic. The mean percentile of the examples selected was 0.65, but this raises to 0.82 if we consider only the stronger signature topic per instance.

A one-sided Mann–Whitney test was conducted on these latter 29 examples, comparing z-normalized "classic instance" topic proportion with that of the genre median, gave a test statistic of $U = 820$ (common language effect size 98%, $p < .01$), supporting the hypothesis that these examples were both consistently and significantly above the genre median in at least one signature topic.

## 4 | DISCUSSION

Caution must of course be exercised in progressing from the characteristics and properties of a probabilistic model to its interpretation (D. Blei, 2012a). In the present case, this should include the recognition that we are taking a textual representation of a work—the plot summary—as a proxy for the work itself, and ignoring many other aesthetic and multimodal aspects. The summary itself may be emphasizing only some scenes or storylines within a film and be dependent on the vocabulary and style of its author or authors.

### 4.1 | Genre characteristics

Despite these significant caveats, we have demonstrated that certain genres can be reliably recognized from their text-based topic distribution, echoing work on both film and literature genre prediction (Ertugrul & Karagoz, 2018; Underwood, 2016b). This works best for those genres with the more characteristic motifs, though even more diffuse genres (e.g., suspense, action/adventure) can still be quite accurately classified. Genre distinctiveness typically relies on one or two of our 70 topics being disproportionately present in the summary for that kind of work.

While these signature topics are important for identification, accompanying topics help to give a more rounded view of a genre and enable the more nuanced

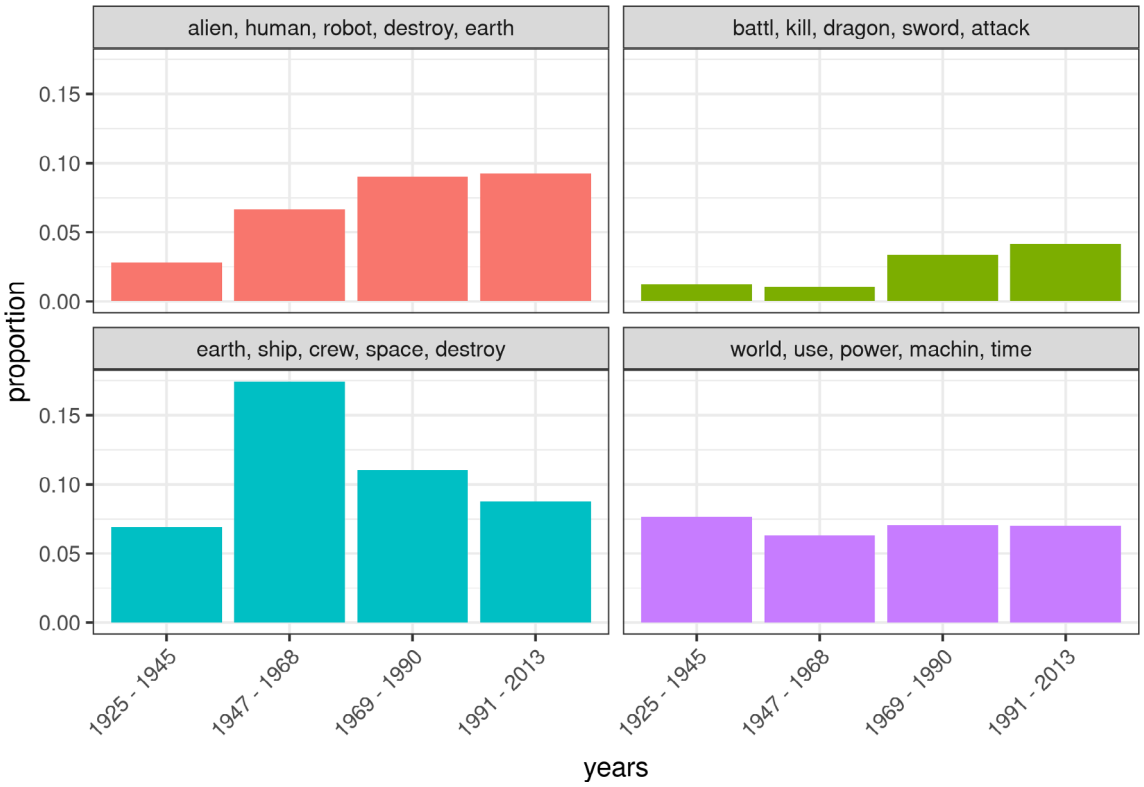



FIGURE 3 Science fiction—Topic proportions over time

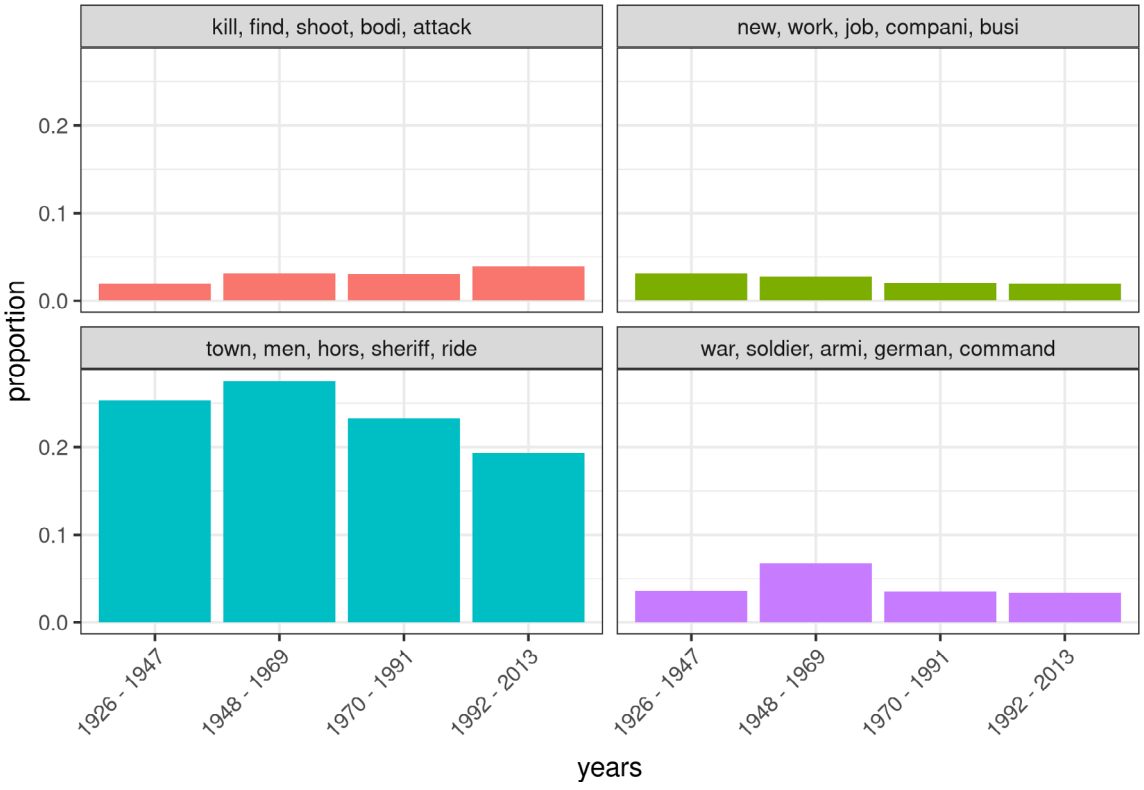



FIGURE 4 Westerns—Topic proportions over time

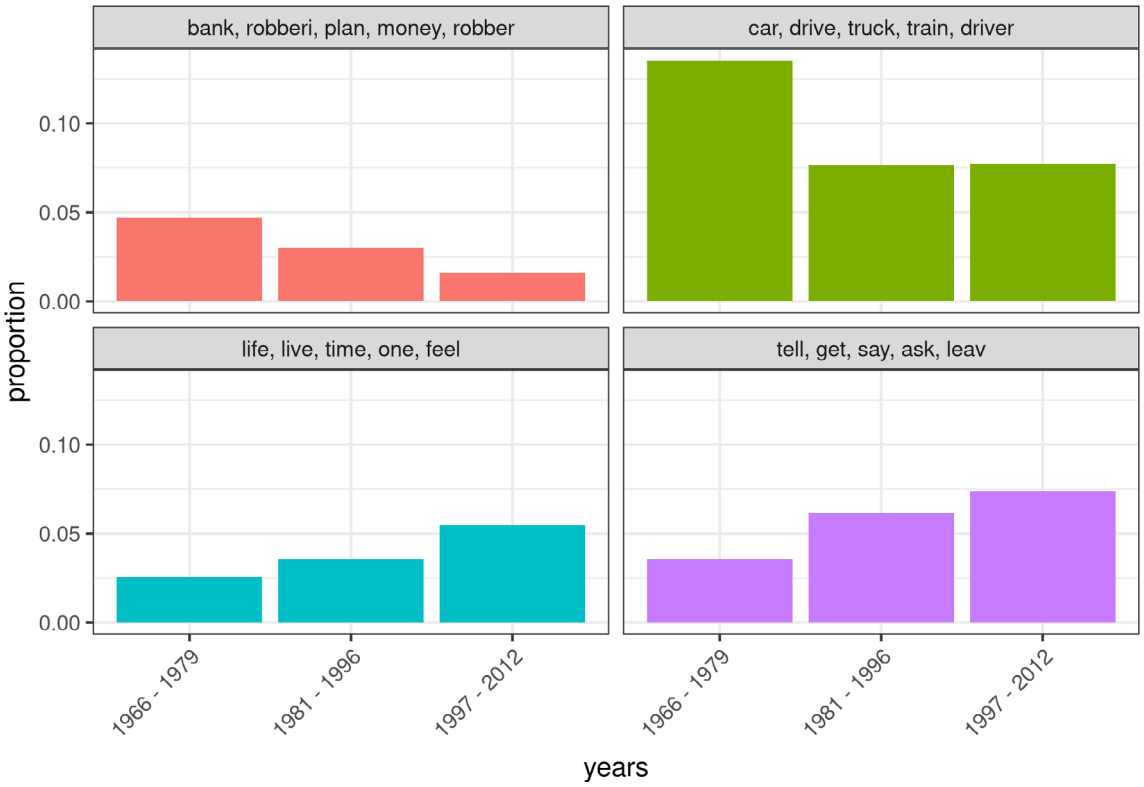



FIGURE 5 Road movies—Topic proportions over time

view of their variance and scope. This somewhat corresponds to the genre definition approach of Altman (1986), who offered a semantic/syntactic distinction in film theorizing, with semantics denoting common traits but the syntax or overall structure being equally important to the genre definition. Thus, common semantic components may belong to different genres based on how they are contextualized or emphasized:

"The distinction between the semantic and the syntactic, in the way I have defined it here, thus corresponds to a distinction between the primary, linguistic elements of which all texts are made and the secondary, textual meanings that are sometimes constructed by virtue of the syntactic bonds established between primary elements." (Altman, 1986, p. 38).

By Altman's lights the individual topics in the present study are identifying semantic elements and the estimated topic composition of overall works is capturing something of the syntax (though not necessarily any sequencing, as topics are aggregated together in our analysis).

## 4.2 | Genres over time

The analysis of topic composition between gross time periods revealed some patterns of interest for the sample genres. First, signature topics tended to be dominant throughout the three or four periods studied. At the same time, there was an overall reduction of these over the 50–80 year window studied. So it seemed that the summaries progressively deemphasized the genre motifs. This may indicate that the works themselves have progressed from a relatively two-dimensional, strongly formulaic period, though to the stage where the genre identification remains important but is also used as a background for the admixture of more diverse elements. If so, we could take it as evidence for the periodic theories of genre histories, with the peak topic prevalence corresponding to the classical/canonization stage. However, there are alternative explanations which should also be explored: for example, have summary styles changed across the board in the periods studied? This may certainly be a factor. While it is unlikely that word meaning itself has changed significantly in the period studied (a concern noted by Schmidt, 2012), it might be that more of the summaries are written from a "fan" perspective than a "studio" perspective, and correspondingly less terse and more detailed or discursive. This indeed may be part of the picture, as there was a weak but significant correlation (0.12) between summary length and release year. A second challenge is knowing if summaries are actually contemporaneous with the films' releases. It may be that those for older works in the corpus have been,

TABLE 11 ANOVA of road movie topic trends

| | *df* | Sum Sq | Mean Sq | *F* value | Pr (>F) |
|---|---|---|---|---|---|
| life, live, time, one, feel | 1 | 5.30 | 5.30 | 9.78 | .00 |
| tell, get, say, ask, leav | 1 | 5.21 | 5.21 | 9.62 | .00 |
| bank, robberi, plan, money, robber | 1 | 2.25 | 2.25 | 4.16 | 0.04 |
| town, men, hors, sheriff, ride | 1 | 2.20 | 2.20 | 4.05 | 0.05 |
| woman, man, old, meet, name | 1 | 1.83 | 1.83 | 3.39 | 0.07 |
| film, scene, peopl, shown, includ | 1 | 1.53 | 1.53 | 2.82 | 0.09 |
| car, drive, truck, train, driver | 1 | 0.73 | 0.73 | 1.34 | 0.25 |
| sex, relationship, sexual, friend, apart | 1 | 0.71 | 0.71 | 1.31 | 0.25 |
| new, work, job, compani, busi | 1 | 0.01 | 0.01 | 0.02 | 0.89 |
| father, mother, sister, daughter, home | 1 | 0.01 | 0.01 | 0.01 | 0.92 |
| Residuals | 212 | 114.83 | 0.54 | NA | NA |

Abbreviation: ANOVA, analysis of variance.

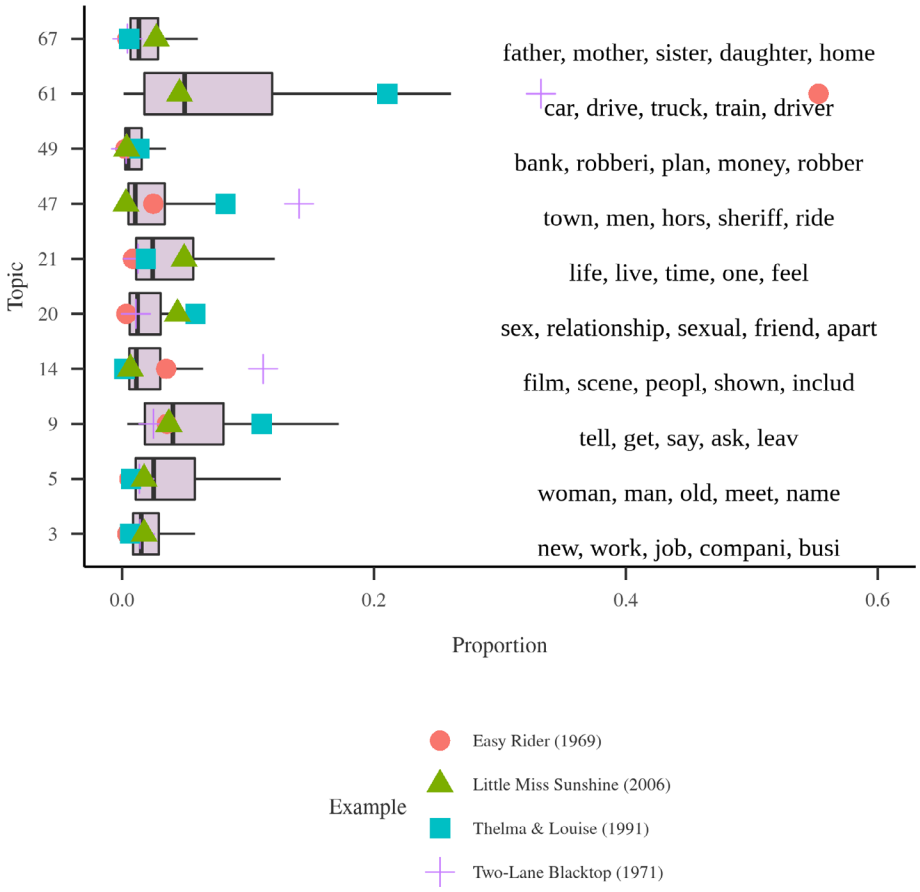



FIGURE 6 Topic proportions for westerns with highlighted examples

retrospectively, applied, meaning that the language and values applied are not of the time.

A further interesting observation is the changes in topical composition between periods for the genres investigated. These seemed to indicate a change of emphasis in motific terms in the case of science fiction, or the increasing emphasis of secondary themes in the case of road movies with the self-actualization aspects becoming stronger. In the latter case, it is again hard to disentangle the actual movie contents from the relative emphasis placed by the summary author. However, there seems to be some corroboration of the "inward" turn in science fiction in the former, with stories mirroring colonial idealism turning to pessimism (Higgins, 2010).

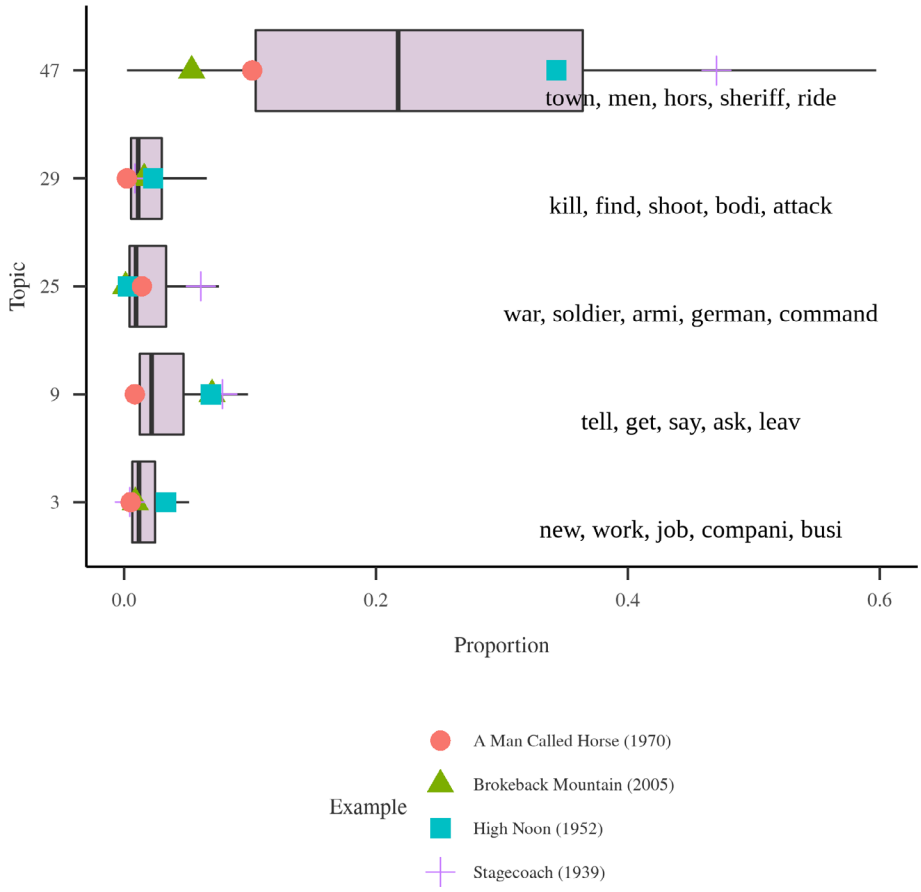



**FIGURE 7** Topic proportions for road movies with highlighted examples

### 4.3 | Canonicity

When looking at the estimated topic proportions of canonical films compared to the overall distribution within the genre, we see that they are frequently either outliers or at the top end of the distribution for signature topics (e.g., road and car for road movies). It should be noted that, for the purposes of our exercise, we conflated "classic" with "canonical," as the former was more accessible to selection. However, the two concepts are closely related. This finding has some interesting potential implications both for the idea of the canon and for prototype theory in classification. It seems to refute the idea of a "canonical center" (Higgins, 2010)—those examples actually at the center (at least in content prevalence terms) are less likely to be deemed canonical. It also casts doubt on whether, for movies, prototypical examples are really representative. They instead seem to be almost caricature, memorable in their extreme focus on the signature motif. The canon seems to function as a snow marker, with many less extreme genre examples less visible or below the surface.

This work has demonstrated topic modeling's potential in classification and discovery, suggesting possible subgenres, subjects or content-based facets. We have also demonstrated potential for discovery through similarity. While we have looked at similarity to modal topic proportions, this could also be done at the level of individual items. Through mapping the genre landscape, we also see some of the problems of genre in differentiating a corpus of works. The example of "World cinema," which tends to deemphasize the range of genres it encompasses, is a case in point. As Dimock (2007) notes of world literature as a genre, this is more of an accident of origin, a virtual category rather than something informational.

## 5 | CONCLUSIONS AND FUTURE RESEARCH

The topic modeling of plot summaries can offer much in understanding how language, themes and genre interrelate. We have seen that the human-labeled genres are often recoverable from the unsupervised machine-detected patterns within the text. These composite thematic distributions enable the study of genre across time, the positioning of examples within a genre category and the identification of candidate items of interest. Experiments suggest that for some genres, signature topics change in emphasis over time. Secondary topics may

become more or less important, indicating a shift in what a situated public perceive as a genre family. They also reveal that canonical examples may not be central in a distributional sense, but more like peaks in an undulating scene.

Additional work might seek to further validate the findings presented here, using other datasets. It will be interesting, for instance, to see if similar patterns are found by topic modeling movie scripts as done in other studies which have focused on prediction (Chao & Sirmorya, 2016). We might expect these to be more variable, however, as the extent of expository narrative will vary between productions. More ambitious, though potentially rewarding, could be to combine textual with visual features, as lighting and motion to give a more holistic approach to genre representation (e.g., Deldjoo, 2016). In terms of genre prediction, while we have shown that topic features can be good predictors of certain genres for the plot summary dataset, it will be instructive to undertake sensitivity analysis with topic models using different topic numbers and hyperparameter values, in order to see how term-topic and topic-document complexity affects prediction. It might be, for instance, that a richer model could work better in detecting more subtle differences in fuzzier genres such as comedy, adventure or romance.

In sum, this article supports a view of genre for film as a multidimensional landscape (Dimock, 2007) where genre is topological rather than typological (Martin & Rose, 2008). This is consistent with an ecological view of genre and the concept of uptake, where a work can be seen as a multigenre set of possibilities, whose selection and relative emphasis is due to the social setting and particular biases of its receivers (institutions and individuals) (Bawarshi & Reiff, 2010).

Our analysis and approach have both theoretical and practical implications. In terms of genre theory, it will lend quantitative evidence to qualitative and critical studies of movie genre composition and dynamics. Practically, this work has potential application to automated genre labeling and to movie discovery through content-based recommendation in terms of both similarity and contrast.

## AUTHOR CONTRIBUTIONS

**Paul Matthews**: Conceptualization, writing - original draft preparation, writing - review and editing. **Kathrina Glitre**: Writing - review.

## ORCID

*Paul Matthews* https://orcid.org/0000-0003-1021-2683

## REFERENCES

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? (and how to fix it using search-based software engineering). *Information and Software Technology*, *98* (February), 74–88. https://doi.org/10.1016/j.infsof.2018.02.005

Altman, R. (1986). A semantic/syntactic approach to film genre. In *Film genre reader III* (pp. 27–41). Austin: University of Texas Press.

Bamman, D., O'Connor, B., & Smith, N. (2014). CMU movie summary corpus. Retrieved from http://www.cs.cmu.edu/~ark/personas/

Bamman, D., O'Connor, B., & Smith, N. A. (2013). *Learning latent personas of film characters*. In Proceedings of the 51st annual meeting ofthe association for computational linguistics (pp. 352–361).

Bawarshi, A. S., & Reiff, M. J. (2010). *Genre: An introduction to history, theory, research, and pedagogy*. Parlor Press Retrieved from https://wac.colostate.edu/books/referenceguides/bawarshi-reiff/

Blei, D. (2012a). Topic modeling and digital humanities. *Journal of Digital Humanities*, *2*(1) Retrieved from http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Blei, D. M. (2012b). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*(2), 634–634. https://doi.org/10.1214/07-aoas136

Chao, B., & Sirmorya, A. (2016). Automated movie genre classification with LDA-based topic modeling. *International Journal of Computer Applications*, *145*(13), 1–5. https://doi.org/10.5120/ijca2016910822

Contributors, W. (2020a). Road movie—Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Road/movie

Contributors, W. (2020b). Science fiction—Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Science/fiction

Contributors, W. (2020c). Western (genre)—Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Western/(genre)

de Esteban, M. (2012). ReGenreNation: The revision of genres at the British film institute. *Catalogue & Index*, *166*, 28–29.

Deldjoo, Y. (2016). Recommending movies based on mise-en-scene design. In *CHI 2016* (pp. 1540–1547). ACM.

Dimock, W. C. (2007). Introduction: Genres as fields of knowledge. *PMLA*, *122*(5), 1377–1387. https://doi.org/10.1632/pmla.2007.122.5.1377

Edelmann, A., & Mohr, J. W. (2018). Formal studies of culture: Issues, challenges, and current trends. *Poetics*, *68*(May), 1–9. https://doi.org/10.1016/j.poetic.2018.05.003

Ertugrul, A. M., & Karagoz, P. (2018). *Movie genre classification from plot summaries using bidirectional LSTM*. Proceedings of the 12th IEEE International Conference on Semantic Computing, ICSC 2018. 248–251. https://doi.org/10.1109/ICSC.2018.00043

Eve, M. P. (2019). *Close reading with computers: textual scholarship, computational formalism, and David Mitchell's Cloud atlas* (p. 251). Stanford University Press.

Evnine, S. J. (2015). "But is it science fiction?": Science fiction and a theory of genre. *Midwest Studies in Philosophy*, *39*(1), 1–28. https://doi.org/10.1111/misp.12037

Furstenau. (1995). *Film genre: The pragmatics of classification*. (PhD thesis). University of Alberta.

Gazan, R. (2015). *First-mover advantage in a social Q&A community*. 2015 48th Hawaii International Conference on System Sciences, 1616–1623. https://doi.org/10.1109/HICSS.2015.195

Geeraerts, D. (1989). Prospects and problems of prototype theory. *Linguistics*, *27*(4), 587–612. https://doi.org/10.17684/i4a53en

Geraghty, L., & Jancovich, M. (2008). Generic canons. In L. Geraghty & M. Jancovich (Eds.), *The shifting definitions of genre: Essays on labeling films, television shows and media* (pp. 1–12). McFarland, CA.

Giaquinto, R., & Banerjee, A. (2018). DAPPER: Scaling dynamic author persona topic model to billion word corpora, 2018 IEEE International Conference on Data Mining (ICDM), pp. 971–976, https://doi.org/10.1109/ICDM.2018.00120

Grant, B. K. (2007). *Film genre: From iconography to ideology* (p. 131). Wallflower.

Henrichs, A. (2019). Deforming Shakespeare's sonnets: Topic models as poems. *Criticism*, *61*(3), 387–412. https://doi.org/10.13110/criticism.61.3.0412

Higgins, D. (2010). *The Inward Urge: 1960s Science Fiction and Imperialism* (PhD thesis No. April). IN.

Hurault-Paupe, A. (2015). The paradoxes of cinematic movement: Is the road movie a static genre? *Miranda*, *10*. https://doi.org/10.4000/miranda.6257

Lakoff, G. (1987). Women, fire, and dangerous things: What categories reveal about the mind. The University of Chicago Press https://doi.org/10.1016/0004-3702(88)90035-5, *35*, 137, 141

Lawrence, E. (2015). Everything is a recommendation Netflix, Altgenres and the construction of taste. *Knowledge Organization*, *42*(5), 358–364.

MacFarlane, A. (2016). Knowledge organisation and its role in multimedia information retrieval. *Knowledge Organization*, *43*(3), 180–183. https://doi.org/10.5771/0943-7444-2016-3-180

Madrigal, A. (2014). How Netflix reverse-engineered Hollywood—the Atlantic. *The Atlantic* Retrieved from https://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/

Martin, J. R., & Rose, D. (2008). *Genre relations: Mapping culture*. Equinox.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). *Optimizing semantic coherence in topic models*. EMNLP 2011—Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, (2), 262–272.

Moretti, F. (2005). *Graphs, maps, trees: Abstract models for literary history*. Verso.

Neale, S. (2007). Questions of genre. In B. K. Grant (Ed.), *Film genre reader III* (p. 636). University of Texas Press.

Patil, C. (2019). Kaggle datasets: Indian_names. Retrieved from https://www.kaggle.com/chaitanyapatil7/indian-names

Roberts, M., Stewart, B. M., Tingley, D., & Airoldi, E. (2013). *The structural topic model and applied social science*. In *NIPS 2013 Workshop on Topic Models: Computation, Application, and Evaluation*.

Schmidt, B. M. (2012). Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities*, *2*(1) Retrieved from http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/

Schmidt, B. M. (2016). Do digital humanists need to understand algorithms? Retrieved from https://dhdebates.gc.cuny.edu/read/untitled/section/557c453b-4abb-48ce-8c38-a77e24d3f0bd{/\#}ch48

Spohrer, E. (2009). Not a gay cowboy movie? Brokeback mountain and the importance of genre. *Journal of Popular Film and Television*, *37*(1), 26–33. https://doi.org/10.3200/JPFT.37.1.26-33

Tudor, A. (1976). Genre and critical methodology. In B. Nichols (Ed.), *Movies and methods: An anthology* (pp. 118–126). University of California Press.

Underwood, T. (2016a). Genre theory and historicism. *Journal of Cultural Analytics*, *1*(1), 8–13. https://doi.org/10.22148/16.008

Underwood, T. (2016b). The life cycles of genres. *Journal of Cultural Analytics*, *70*(1984), 151–167. https://doi.org/10.22148/16.005

Wickham, H. (2019). GitHub—hadley/babynames: An R package containing US baby names from the SSA. Retrieved from https://github.com/hadley/babynames

**How to cite this article:** Matthews, P., & Glitre, K. (2021). Genre analysis of movies using a topic model of plot summaries. *Journal of the Association for Information Science and Technology*, *72*(12), 1511–1527. https://doi.org/10.1002/asi.24525