Rating consistency is consistently underrated: An exploratory analysis of movie-tag rating inconsistency

Denis Kotkov
Department of Computer Science
University of Helsinki
Helsinki, Finland
kotkov.denis.ig@gmail.com

Alan Medlar
Department of Computer Science
University of Helsinki
Helsinki, Finland
alan.j.medlar@helsinki.fi

Umesh Raj Satyal
Faculty of Science and Engineering
Åbo Akademi University
Turku, Finland
satyalumesh@gmail.com

Alexandr Maslov
Faculty of Science and Engineering
Åbo Akademi University
Turku, Finland
maslov314@gmail.com

Mats Neovius
Faculty of Science and Engineering
Åbo Akademi University
Turku, Finland
mats.neovius@abo.fi

Dorota Glowacka
Department of Computer Science
University of Helsinki
Helsinki, Finland
dorota.glowacka@helsinki.fi

ABSTRACT

Content-based and hybrid recommender systems rely on item-tag ratings to make recommendations. An example of an item-tag rating is the degree to which the tag "comedy" applies to the movie "Back to the Future (1985)". Ratings are often generated by human annotators who can be inconsistent with one another. However, many recommender systems take item-tag ratings at face value, assuming them all to be equally valid. In this paper, we investigate the inconsistency of item-tag ratings together with contextual factors that could affect consistency in the movie domain. We conducted semi-structured interviews to identify potential reasons for rating inconsistency. Next, we used these reasons to design a survey, which we ran on Amazon Mechanical Turk. We collected 6,070 ratings from 665 annotators across 142 movies and 80 tags. Our analysis shows that ~45% of ratings are inconsistent with the mode rating for a given movie-tag pair. We found that the single most important factor for rating inconsistency is the annotator's perceived ease of rating, suggesting that annotators are at least tacitly aware of the quality of their own ratings. We also found that subjective tags (e.g. "funny", "boring") are more inconsistent than objective tags (e.g. "robots", "aliens"), and are associated with lower tag familiarity and lower perceived ease of rating.

CCS CONCEPTS

• **Information systems** → **Social tagging**; *Recommender systems*; *Users and interactive retrieval*; • **Human-centered computing** → **User studies**;

KEYWORDS

recommender systems; critiquing recommender systems; item-tag rating; tag relevance; tagging; tag genome; rating inconsistency

ACM Reference Format:
Denis Kotkov, Alan Medlar, Umesh Raj Satyal, Alexandr Maslov, Mats Neovius, and Dorota Glowacka. 2022. Rating consistency is consistently underrated: An exploratory analysis of movie-tag rating inconsistency. In *The 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22), April 25–29, 2022, Virtual Event, .* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3477314.3507270

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SAC '22, April 25–29, 2022, Virtual Event,
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/10.1145/3477314.3507270

1 INTRODUCTION

Many online systems use keywords to concisely describe the content of items, such as books and movies. For example, Netflix [4] uses genre tags to describe movies and television programmes, Quora [5] uses tags to categorise questions, and YouTube [6] allows users to attach hashtags to videos. Throughout this paper, we use the term *tag* to refer to any sequence of words that can be used to describe the content of an item.

Tags can be used to describe content [13, 24], users [26], and to search [14], filter [15] and recommend [16] items. Item-tag ratings, however, where tags are scored by the degree to which they apply to a given item, are most often used in various types of recommender systems. Content-based and hybrid recommender systems use item-tag ratings to make recommendations [9, 12, 31, 38]. Tag recommender systems suggest tags for user-item pairs [25, 27, 29]. Critiquing recommender systems rely on rich item descriptions, including tags, to recommend items interactively. For example, Movie Tuner uses movie-tag ratings from Tag Genome to allow users to adjust recommendations using qualitative statements, such as "more comedy" or "less magic" [38].

Item-tag ratings can be generated based on item content [10, 32, 33] or entered by annotators [11, 14, 21, 22, 38]. In this paper, we use the term *annotator* to refer to any individual who enters item-tag ratings. Annotators can be domain experts, who follow extensive instructions to enter consistent ratings [11], or non-expert users of the recommender system itself who may have limited domain knowledge and are not given detailed annotation instructions [14,

21, 22, 38]. In this paper, we focus on the latter scenario due to its popularity in the literature [14, 21, 22, 38].

Recommender systems that use either content-based [10, 32, 33] or manually annotated [11, 14, 38] item-tag ratings assume they are reliable, which may not be the case. Indeed, a recent study by Kotkov et al. highlighted the presence of item-tag rating inconsistency in Tag Genome, with ratings for subjective tags, such as "funny" and "overrated", being more inconsistent than objective tags, such as "space" and "time travel" [21]. However, the conclusions of this study may have been biased by Tag Genome's data collection method, which resulted in long-tailed distributions of movies and tags [38]. In this paper, we systematically investigate the inconsistency of item-tag ratings in the movie domain and additionally explore contextual factors, such as how long ago the movie was watched, that could impact the overall quality of annotations. Our research questions are as follows:

RQ1: How inconsistent are movie-tag ratings?
RQ2: Does rating consistency vary between different kinds of tags, e.g. subjective versus objective, and, if so, what factors contribute to these differences?
RQ3: What contextual factors are associated with movie-tag rating inconsistency?

To answer these research questions, we conducted semi-structured interviews with annotators to identify potential reasons why they might assign inconsistent ratings to item-tag pairs. Based on these findings, we designed a survey that asked annotators to rate movie-tag pairs and to provide additional information on their ratings. We ran the survey on Amazon Mechanical Turk [1] and collected 6,070 ratings from 665 annotators across 142 movies and 80 tags.

Our analysis of the survey responses found that ~45% of ratings were inconsistent with the mode rating (i.e. the most frequent rating) for a given movie-tag pair. This inconsistency was associated with annotators' perceived ease of rating, with lower ease of rating linked to lower consistency. We replicated the findings of Kotkov et al. [23] showing that subjective tags were rated more inconsistently than objective tags, however, we additionally show that annotators were more familiar with the meaning of objective tags and found them easier to rate compared to subjective tags.

Our findings are of particular interest to researchers and practitioners in recommender systems. First, our results motivate the discussion regarding the assumptions underpinning popular systems, such as Movie Tuner, and datasets, such as Tag Genome [38]. Second, we show that item-tag rating inconsistency is related to item rating inconsistency, which is an important topic in recommender systems [7, 17, 18]. Third, our findings can lead to improvements of content-based and hybrid recommender systems which often rely on item-tag ratings [9, 12, 31, 38].

The contributions of this paper are as follows:

(1) We provide a more detailed analysis of tag rating inconsistencies based on a more fine-grained categorization of tags compared to previous studies [21].
(2) We investigate numerous contextual factors that are potential causes of movie-tag rating inconsistency.
(3) We provide a publicly available dataset of survey responses collected as part of the study[1].

[1]https://github.com/Bionic1251/Rating-consistency-is-consistently-underrated

2 RELATED WORKS

The inconsistency of ratings has been studied extensively for items, but less so for item-tag pairs. Here, we give a brief overview of previous studies.

2.1 Inconsistency of item ratings

Inconsistency of item-tag ratings is related to inconsistency of item ratings. Item ratings usually indicate the degree to which a user enjoyed consuming an item. This topic has been widely studied in recommender systems [7, 17, 18, 30, 35].

The inconsistency of ratings that users provide for items was first studied in the context of recommender systems in 1995 [18]. In the study, users were asked to rate a set of movies on a scale from 1 to 10 and then rerate them six weeks later. Some of the ratings provided by users differed from their earlier ratings. In particular, the Pearson correlation coefficient of sets of ratings given by 19 users at different times was 0.83. The authors of this study suggested that user ratings contain noise which prevents them from ever being predicted perfectly. Indeed, Herlocker et al. noticed that the performance of recommendation algorithms approaches an upper limit, which they termed the *magic barrier*, which refers to the minimum error with which a recommender system can predict user ratings.

Amatriain et al. investigated the magic barrier, identifying several contributory factors [7]: (1) users tend to be more consistent with ratings at the extremes of a scale; (2) the order in which users are asked to provide feedback impacts consistency; and (3) that the time spent on rating items does not impact consistency. In a later study, Amatriain et al. investigated whether user rerating would increase accuracy [8]. They found that rerating items with extreme ratings results in higher accuracy gains than rerating other types of ratings. Furthermore, that rerating items can result in higher accuracy gains than collecting new ratings.

Said et al. proposed a mathematical definition of the magic barrier as the standard deviation of ratings [35]. They estimated the barrier for movie recommendation to be 0.61 on a scale between 0 and 10 with a step size of 0.5. The authors also found that users are more consistent when they rate movies higher than their average rating.

Nguen et al. evaluated different user interfaces in terms of their effect on rating inconsistency [30], introducing an interface that reminds users of items given a similar rating. This interface reduced rating inconsistency, but it also increased cognitive load.

Sergej Sizov demonstrated that users are inconsistent not only over an extended period of time, but also within the same rating session [36]. The author conducted an experiment where he asked 110 participants to rate a set of photos of attractions, with some of the photos shown to participants five times. According to his results, 16% of participants gave the same rating to the photos every time, 50% of participants changed their ratings once and the rest – two or more times.

Said and Bellogín proposed a user coherence measure, which indicates the consistency of user ratings [34]. The measure is based on user ratings and item attributes, such as genre, intended audience and keywords. The authors also showed that user coherence can distinguish between users with higher and lower magic barriers.

Question		Responses	Reason
Q1.	How long ago did you watch [movie]?	Within last 12 months, Between 1 and 5 years, Between 6 and 10 years, More than 10 years, I do not remember	4
Q2.	On a scale from 1 to 5, how strongly does the tag [tag] apply to [movie]?	1-5, Not sure	n/a
Q3.	To what degree do you agree with the statement "it was easy for me to rate the tag [tag] for the movie [movie]"?	Strongly disagree – strongly agree	2
Q4.	On a scale from 1 to 5, how familiar are you with the term [tag]?	1-5	3
Q5.	On a scale from 1 to 5, how often do you watch movies that could be described as [tag]?	1-5	5
Q6.	Please write down three terms or phrases that you associate with [tag].	Free text	3

Table 1: Survey questions, valid responses and reasons for inclusion from Section 3.1.

Jasberg and Sizov proposed a theoretical framework based on metrology for modeling rating inconsistency [19]. The authors conducted an experiment where they asked 67 participants to rate and rerate theatrical trailers, and to provide probability distributions of their ratings. For probability distributions, the authors asked participants to rate the appropriateness of each n-star rating. Both rerating and probability distribution methods resulted in similar distributions of ratings. The authors used the collected data to verify their theoretical framework and showed how rating inconsistency can lead to evaluation errors in recommender systems.

Despite being similar to inconsistency of item-tag ratings, item rating inconsistency has three key differences:

(1) Item ratings indicate the degree to which the user enjoyed consuming the item, while item-tag ratings usually indicate the degree to which a tag applies to an item regardless of whether the user liked it or not (there are exceptions, however, such as tags like "enjoyable").
(2) Some item-tag pairs must have a consensus rating as they describe factual information about items. On the other hand, items need not have a consensus rating as they indicate user opinions.
(3) Item rating inconsistency is related to the inconsistency of individuals' ratings over time. While in this paper, we study inconsistency of item-tag ratings between annotators.

2.2 Inconsistency of item-tag ratings

To the best of our knowledge, only a single study has investigated the consistency of item-tag ratings [21]. The authors conducted an analysis of movie-tag ratings from Tag Genome [38], identifying inconsistent ratings given by different annotators to the same movie-tag pairs. They found that subjective tags, on average, have higher inconsistency than objective ones. However, both subjective and objective tags had moderate inconsistency in general. Although we also investigate inconsistency of movie-tag pairs, our work has a number of differences:

(1) We collect a more balanced dataset than the one used in [21], which suffered from long-tailed distributions of movies and tags [38].
(2) We investigate contextual factors, such as the annotators perceived ease of rating and their familiarity with the meaning of a given tag, to understand whether they influence item-tag rating consistency.

3 SURVEY CONSTRUCTION

We recruited two pairs of participants (two PhD students and two postdoctoral researchers) from our university, and conducted semi-structured interviews with them to investigate why annotators might disagree on tag relevance ratings. The recruited participants work in the areas of computer science (machine learning) (2 participants), medicine (1 participant) and literature (1 participant). They regularly watch movies in streaming services and cinemas, and read books. Prior to the interview, we asked participants to rate several movies and books with tag relevance ratings and to write short definitions for each tag. During the interviews, we asked each pair of participants to go through their ratings together. Where they disagreed, they were asked to either revise their ratings or "agree to disagree" and keep their ratings different.

3.1 Interviews

We took notes during the discussions between the participants in order to identify a list of reasons why their tag ratings differed. These reasons were discussed with each participant afterwards along with the tag definitions they provided. We found the following reasons for disagreement:

(1) Tags can be subjective, e.g. "good movie". The participants disagreed on the ratings due to their differing opinions.
(2) Tags can be more or less difficult to rate ("It is just very difficult to rate the relevance of this tag").
(3) Participants have different understanding of tags ("I assumed that time travel involves the actual travel in time, not when the character recalls their past").
(4) Participants misremember items ("Oh, yes, I forgot about that scene").
(5) Participants have different rating scales depending on past experience ("I have watched more dramatic movies than this, I'd say it is a 4", "I would still keep it a 5").
(6) Participants have not consumed items ("I have not watched it, but I know the plot").

(7) Survey options affect participant answers ("I adjusted my scale while going through other questions").
(8) Participants make mistakes ("I meant to give a different rating. I gave this one by mistake").
(9) Participants change their minds ("Now, when I think more about it, I would change my rating").

We took all these reasons into consideration when designing our survey and when analyzing results (with the exception of reasons 8 and 9).

3.2 Survey design

Table 1 shows the survey questions we created based on the reasons given for annotator disagreement of tag ratings. The survey consisted of four pages. On page 1, annotators identified movies they had already seen and on pages 2 – 4 they answered questions related to movies, tags and movie-tag pairs.

Page 1 showed a list of popular movies and annotators were instructed to select the first 10 movies they had seen. We randomized the ordering of movies to ensure that all movies had an equal chance of being rated and to mitigate the impact of question ordering (reason 7). If annotators had seen fewer than 10 of the available movies, they could still take the survey. We additionally included three fake movies near the top of the list. Annotators were ineligible for the survey if they had not seen any of the movies listed or if they selected any of the fake movies. On page 2, we asked annotators how long ago they watched each movie (Q1 in Table 1). On page 3, we asked annotators to rate movie-tag pairs (Q2) and indicate how easy it was to provide their ratings (Q3). We constructed movie-tag pairs by randomly selecting one of the tags we associate with each movie (see Section 3.3.1). On page 4, we asked annotators about tags. Specifically, annotators were asked to indicate their familiarity with the meaning of different tags (Q4), how often they watch movies related to each tag (Q5), and to provide terms and phrases associated with each tag (Q6).

To help users answer the questions and to allow them to benefit from the anchoring effect [30, 37] (reason 5), we provided high and low scoring examples for each movie-tag pair. We additionally reminded annotators of the content of movies (reason 4) and the definition of tags (reason 3) by including links to the IMDB [2] movie database and Google search for tags throughout the survey. Finally, as some movies were exceptionally popular, we gradually removed movie-tag pairs that had already received over 5 ratings. The lowest number of movies shown to a user on page 1 was 23 (20 real and 3 fake movies) and the highest – 145 (142 real and 3 fake movies).

3.3 Tag and movie selection

Vig et al. [38] conducted a survey in the movie recommender system MovieLens [3]. MovieLens annotators were asked to indicate the degree to which a tag applies to a movie on a 5-point Likert scale (1 corresponds to "not at all" and 5 to "very much"). The annotators could also indicate that they were not sure about the answer (the "not sure" option). Based on the annotator replies and movie metadata, the authors generated the Tag Genome dataset containing relevance scores for movie-tag pairs. The scores are continuous values between 0 and 1 to indicate the degree with which a tag applies to a movie. We used data from the MovieLens survey to select tags for our survey and the Tag Genome dataset to select movies.

3.3.1 Tag selection. The MovieLens survey [38] contains 58,903 ratings provided by 679 annotators to 45,914 movie-tag pairs (5,546 movies and 1,084 tags). We excluded "not sure" ratings, which resulted in 51,163 ratings, 679 users, and 40,013 movie-tag pairs (5,192 movies and 1,084 tags).

A rater categorized all 1,084 tags from the MovieLens survey into four categories: objective, subjective, genre, and miscellaneous (misc.). Objective tags are related to factual information (e.g. "pirates", "James Bond"). Subjective tags are related to opinions, preferences and emotional reactions (e.g. "good plot", "funny"). Genre tags represent movie genres, such as "drama", "action" and "art house", which were considered a separate category due to their importance in describing movies. Finally, misc includes miscellaneous tags that do not fall neatly into any of the three categories. For example, broad concepts, such as "justice", that could describe an event in the movie or the feeling that the movie evokes.

For each category, we selected 20 of the most popular tags from the MovieLens survey, selecting 80 tags in all. We excluded tags that were too specific, such as names of actors, characters, directors and studios. To ensure the tags were categorised correctly, a second rater recategorised all 80 tags. The inter-rater agreement was 0.767 ($P < 10^{-16}$), indicating substantial agreement [28]. Both raters resolved disagreements, which resulted in the following distribution of tags per category: objective – 21; subjective – 22; genres – 18; misc – 19. The full list of tags is available in Table 5.

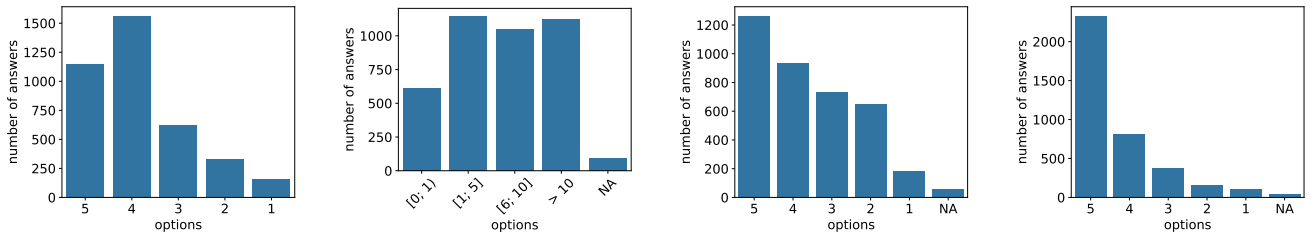
3.3.2 Movie selection. We used the Tag Genome dataset to select movies for each of the 80 tags. For each tag, we ranked all the movies by their tag relevance and extracted movies with low (bottom 25%), medium (middle 25%) and high (top 25%) relevance scores. From the movies with low and high tag relevance, the most popular movies were used as rating anchors in the survey, i.e. as examples of movies with low/high tag relevance. For rating, we extracted the two most popular movies with low, medium, and high tag relevance that had not already been used as rating anchors. This process identified 142 unique movies for rating. We allowed each movie to be associated with multiple tags.

4 DATASET CHARACTERISTICS

We submitted our survey to Amazon Mechanical Turk and received 6,070 ratings from 665 annotators over the course of six days. We rejected surveys from annotators who: (1) did not complete the survey, (2) selected any of the fake movies, (3) selected no movies, or (4) provided random comments and/or tag associations. Afterwards, we were left with 4,022 ratings from 452 annotators.

4.1 Overview

Figure 1 shows the distributions of responses to survey questions that correspond to rating factors. Most annotators agreed with the statement that it was easy to rate movies in terms of a given tag, with over 2/3 of annotators either agreeing or strongly agreeing (Figure 1(a)). Annotators frequently watch movies associated with the tags given (Figure 1(c)) and more than 60% of annotators stated that they



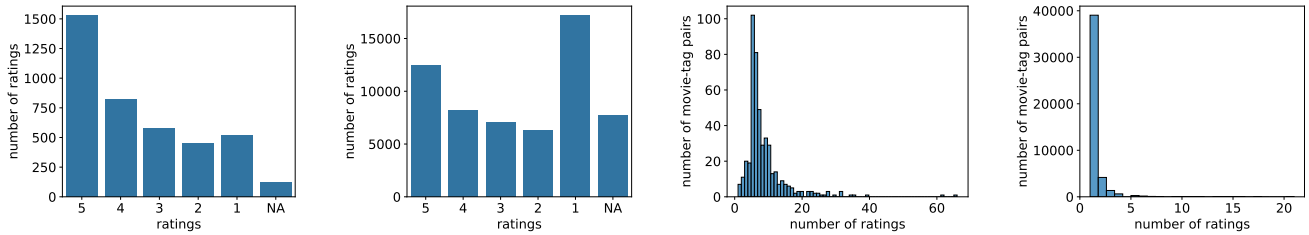
(a) Distribution of answers to the question regarding ease of rating (Q3).

(b) Distribution of answers to the question regarding how long ago the annotator watched the movie (Q1).

(c) Distribution of answers to the question regarding how often the annotator watches movies with the given tag (Q5).

(d) Distribution of answers to the question regarding familiarity of the annotator with the tag (Q4).

Figure 1: Overview of rating factors. The "NA" option corresponds to "I don't remember" or "Not sure", while in the question related to how easy it was to provide each rating, numbers 1-5 correspond to the 5-point Likert scale.



(a) Distribution of ratings in our survey.

(b) Distribution of ratings in the MovieLens survey.

(c) Distribution of ratings per movie-tag pair in our survey.

(d) Distribution of ratings per movie-tag pair in the MovieLens survey.

Figure 2: Comparison between our survey and the MovieLens survey. The "NA" option corresponds to "Not sure".

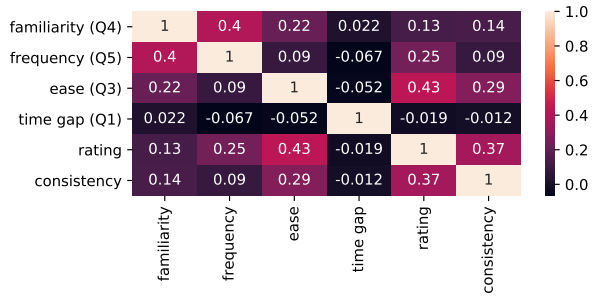


Figure 3: Spearman correlation heatmap between survey questions, ratings and rating consistency. The consistency variable is 1 if the rating corresponds to the mode rating for a given movie-tag pair and 0 otherwise.

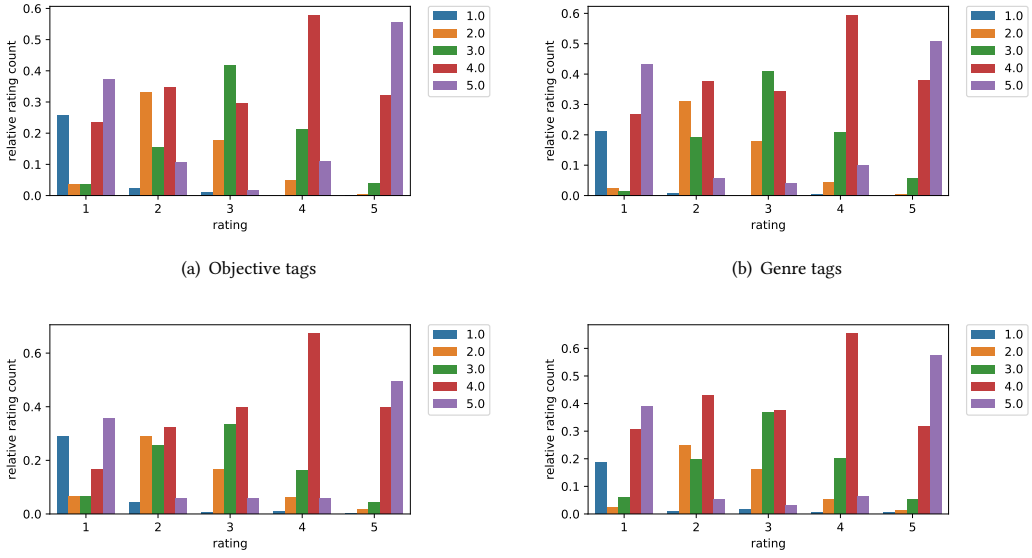
are highly familiar with the meaning of a given tag (Figure 1(d)). In terms of how long ago annotators watched the movies they rated, a similar number of movies were watched in the last 1–5 years, 6–10 years and more than 10 years ago. In the last 12 months, however, annotators watched more movies, on average, than in any preceding period (Figure 1(b)).

4.2 Dataset comparison

Even though our survey and the MovieLens survey [38] included an equal number of movies from the top, middle and bottom of the rating scales for each tag, the movies in our survey resulted in more selected tags and were easier to tag by annotators (i.e. we had far fewer "not sure" ratings) than in the MovieLens survey (Figures 2(a) and 2(b)), which may be due to two reasons. First, to select movies, we used relevance scores calculated based on the whole MovieLens survey, while the authors of the MovieLens survey used only a small fraction of these scores. Second, in situations where the scores were non-precise or annotators made mistakes, a tag was more likely to apply to a movie, because we selected only the most popular tags and movies from the MovieLens survey.

Figure 2(c) shows the distribution of the number of ratings received by each movie-tag pair. Our goal was to collect at least five ratings per movie-tag pair and we achieved this for 413 movie-tag pairs (excluding "not sure" ratings). The most frequently occurring number of ratings was 5 (102 movie-tags) and the maximum was 66. A majority of annotators (78%) were able to provide ratings for ten movies. In the MovieLens survey (Figure 2(d)), most movie-tag pairs had 1 rating (39,047 movie-tag pairs out of 45,914) and the maximum was 21.

Our dataset has a similar distribution of average tag ratings to the MovieLens survey. The Spearman's rank correlation for the mean ratings of the 60 movie-tag pairs found in both datasets with



(a) Objective tags

(b) Genre tags

(c) Subjective tags

(d) Misc tags

Figure 4: Distribution of ease of rating scores for each rating value (1-5) for each of the four tag categories. The bars are proportional and sum to 1 for each rating value.

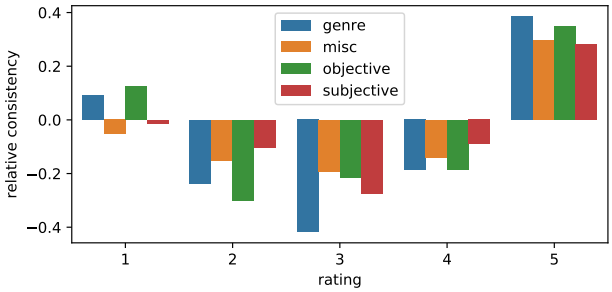


Figure 5: Proportions of consistent ratings per rating value and tag category. The relative consistency has 0.5 subtracted: negative values indicate that a majority of ratings are inconsistent and positive values indicate a majority of ratings are consistent with the mode rating for a given movie-tag pair.

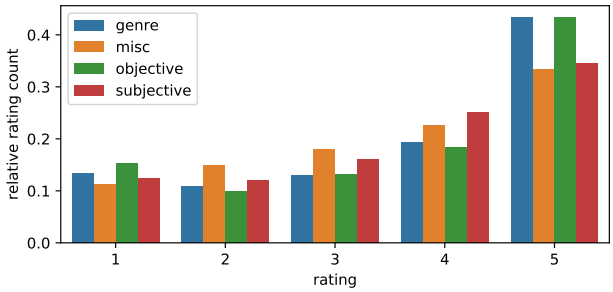


Figure 6: Distribution of ratings based on rating value and tag category. The relative counts are proportional to the number of ratings in a given tag category, i.e. the bars sum to 1 for each tag category.

at least 5 ratings was 0.7 ($P < 10^{-9}$), indicating a strong correlation. However, both datasets clearly contain errors. For example, in the MovieLens survey, the movie-tag pair "Austin Powers: The Spy Who Shagged Me (1999)" – "satire" has an average score of 2.4, despite the movie being a satire of James Bond. Meanwhile in our survey, "Stargate (1994)" – "time travel" has an average score of 4.3 points, despite the movie taking place on another planet and not in ancient Egypt.

4.3 Tag categories and rating patterns

In this section, we focus on rating patterns and rating consistency for different tag categories. For this analysis, we removed the "not sure" ratings, which resulted in 3902 movie-tag ratings. We define a rating as being consistent if it corresponds to the mode (most frequent) rating for a given movie-tag pair. Overall, ~55% of the ratings we collected were consistent (RQ1). Our dataset is slightly biased towards objective tags. It contains 1228 ratings for objective movie-tag pairs, 975 – subjective, 944 – genre and 755 – misc. The following observations can be made about the dataset:

(1) **Annotators found rating easier when tags applied to a movie strongly or very strongly.** While ease of rating had only a moderate correlation with the actual ratings themselves (Figure 3), annotators found it easier to rate movie-tag pairs when they believed that a tag applied to a movie either strongly (4) or very strongly (5) (Figure 4). Furthermore,

Table 2: Category terms from six mixed-effect logistic regression models comparing pairwise rating consistency between tag categories. Significance codes: "*" < 0.002.

Model	Coefficients
objective (1) and subjective (0)	**0.306***
objective (1) and misc (0)	0.258
genre (1) and subjective (0)	0.237
genre (1) and misc (0)	0.194
objective (1) and genre (0)	0.073
misc (1) and subjective (0)	0.045

when a tag should not be applied to a given movie, the ease of rating was polarized, having a similar number of ratings at both ends of the ease of rating scale. These trends were the same across all four tag categories.

(2) **Annotators found rating easier when they gave more consistent ratings.** Ease of rating and rating consistency have a low, but positive, correlation (Figure 3).

(3) **Annotators are more consistent at the extremes of the rating scale.** Figure 5 suggests that annotators are mostly consistent when they indicate that a tag strongly applies to a movie (we observed the same trend in the MovieLens survey data). For genre and objective tags, annotators were also mostly consistent when that tag did not apply to the movie, i.e. a rating of 1. These findings replicate prior studies [7].

(4) **Objective and genre tags have the highest proportion of extreme ratings.** Figure 6 indicates that annotators rate movies-tag pairs with 1 and 5 more frequently for objective and genre tags than for other categories.

(5) **Annotators are more familiar with tags related to the movies they watch.** The correlation matrix (Figure 3) suggests that tag familiarity has a moderate positive correlation with watching frequency. This finding is to be expected because watching movies with a particular tag is likely to make the annotator more familiar with that tag.

5 ANALYSIS

To answer our remaining research questions, we applied regression analysis to different variables and subsets of the collected data. To answer RQ2, we used logistic regression to predict rating inconsistency between tag categories and ordinal regression to predict contextual factors with tag categories, and to answer RQ3, we used logistic regression to predict rating inconsistency on the basis of contextual factors and tag categories. We used mixed-effect regression models to take into account repeated measures for users. To avoid false discoveries, we applied the Bonferroni correction to our significance level. We conducted 19 statistical tests, which corrects our significance level as follows: $0.05/19 = 0.002$.

Table 3: Category terms from four cumulative link mixed-effect ordinal regression models comparing objective and subjective tag categories using each survey question. Significance codes: "*" < 0.002, "*" < 0.00005.**

Dependent variables	Coefficients
familiarity (Q4)	**1.076*****
ease of rating (Q3)	**0.325***
watching frequency (Q5)	0.171
watching time gap (Q1)	0.080

5.1 Subjective tags are more inconsistent than objective tags

To understand whether there was a difference in movie-tag rating consistency between tag categories, we fitted six mixed-effect logistic regression models. Each model compared movie-tag ratings from two categories, e.g. objective and subjective tags. In each logistic regression model, the dependent variable is whether each movie-tag rating was consistent with the mode (most frequent) rating for a given movie-tag pair. The independent variable is a binary variable indicating which category the tag belonged to. To account for repeated measures, we included a random intercept for each user (i.e. in an R-style formula: equals_mode ~ category + (1|user_id)).

Table 2 shows the coefficients for the category term from each of the six logistic regression models, which were statistically significant in only a single model: objective and subjective. These findings show that annotators rate subjective tags more inconsistently than objective tags (RQ2), which corresponds to findings from [21]. The odds of a rating being consistent for an objective tag are 36% higher on average than for a subjective tag. Although genre movie-tag pairs have a similar distribution of ratings compared to objective tags (Figures 5 and 6), there was no significant difference between genre and any other tag category after correction for multiple comparisons. Before multiple comparison correction, however, there was significant differences between genre and subjective, genre and misc, and objective and misc categories (in each case, the first category was more consistent than the second), suggesting that these differences might have been significant had we collected more data.

5.2 Familiarity and ease of rating correlate with rating inconsistency between subjective and objective tags

Next, we wanted to understand what factors might influence the difference in ratings consistency between objective and subjective tags using questions 1, 3, 4 and 5 from our survey. We fitted four cumulative link mixed-effect ordinal regression models, where each model corresponded to a different survey question. In each ordinal regression model, the dependent variable is an ordinal response scale for one of the survey questions and the independent variable is a binary variable indicating whether the tag was from the objective tag category or the subjective category. To account for repeated measures, we included a random intercept for each user (i.e. in an R-style formula: response ~ category + (1|user_id)).

Table 3 shows the coefficients for the category term from each of the four ordinal regression models, which were statistically significant in the model for Q3 (ease of rating) and Q4 (tag familiarity). These findings show that annotators are more familiar with objective tags than subjective tags and, furthermore, that they find objective tags easier to rate (RQ2). For objective tags, the odds that annotators answer one point higher on the response scale compared to subjective tags are 193% and 38% higher for tag familiarity (Q4) and ease of rating (Q3), respectively. There are two possible reasons why subjective tags could be more difficult to understand and rate than objective tags: users may tend to choose more ambiguous tags to express opinions than to indicate facts about movies (recall we selected the most popular tags from each category during survey construction (section 3.3.1)). Second, subjective feelings might simply be harder to express with tags and there are no simpler alternatives.

The fact that objective tags have higher ease of rating seems to be unrelated to annotators' familiarity with these tags as there is only a weak correlation (Figure 3, for the subsets of subjective and objective tags these correlations are very similar). The difference in ease of rating could be attributed to objective tags having a higher fraction of high ratings (Figure 6) than subjective tags, as ease of rating has a moderate correlation with ratings (Figure 3).

5.3 Rating consistency is correlated with ease of rating

Lastly, we wanted to understand what factors influence ratings inconsistency in general. We fitted a mixed-effect logistic regression model, where the dependent variable is whether the rating corresponds to the mode rating for a given movie-tag pair and the independent variables are the tag categories (with objective as the baseline category) and questions 1, 3, 4 and 5 from the survey (i.e. in an R-style formula: equals_mode ~ category + familiarity + watching_time_gap + watching_frequency + ease_of_rating + (1|user_id)).

Table 4 shows the coefficients for the fixed effects from the logistic regression model, none of which were significant with the exception of ease of rating (Q3). This finding shows that when annotators found it easier to rate a movie-tag pair, they were more likely to give a consistent rating (RQ3). In particular, we see a 59% increase in the odds of consistent rating for a one unit increase in ease of rating. Other factors might also be important for rating consistency, but we do not have enough evidence to make this claim. For example, the p values for tag category are ~0.02, while for tag familiarity it is ~0.003, which were statistically significant prior to Bonferroni correction.

6 DISCUSSION AND LIMITATIONS

Below, we discuss the implications of and insights from the process of rating movie-tag pairs, as well as the limitations of our study.

6.1 Ease of rating

The fact that ease of rating is the strongest predictor of rating inconsistency suggests that we can, to a certain degree, trust annotators' judgement regarding the quality of their ratings. This finding could have implications on the tag collection process. For example, if a few annotators provide their ratings and indicate that the ratings were easy to give, then this data should be sufficient to draw strong conclusions, with further collection of ratings halted.

Table 4: The coefficients for fixed effects from a mixed-effect logistic regression model containing all survey data, where the dependent variable is whether the movie-tag rating corresponds to the mode rating for a given movie-tag pair. Significance codes: "*" < 0.00005.**

Independent variables	Coefficients
misc (1) or subjective (0)	0.001
genre (1) or subjective (0)	0.152
objective (1) or subjective (0)	0.215
familiarity (Q4)	0.122
watching time gap (Q1)	-0.020
watching frequency (Q5)	0.049
ease of rating (Q3)	**0.466*****

Improving ease of rating might reduce rating inconsistency. However, this topic requires additional investigation as to why annotators find movie-tag pairs difficult to rate. One of the factors affecting consistency or ease of rating could be the rating scales themselves [19]. For example, adding additional explanations on the scale or reducing the number of options could make rating easier. It is possible that ease of rating cannot be significantly improved, however, as this factor has negligible correlation with factors related to only tags or only movies (Figure 3) and this is the only factor correlated with the combination of a movie and a tag. This might suggest that the movie provides the context for the tag and vice versa. For example, even if the annotator is not highly familiar with the tag "stylized", when the tag appears in a combination with the movie "Sin City (2005)", it becomes easy to rate, while with other movies this rating is difficult to provide.

6.2 Tag categories

It seems that one of the key reasons why objective tags have higher rating consistency than subjective ones is the difficulty of annotators to mark the absence of tags in movies. It is possible that for many subjective tags, such as "surreal" or "inspirational", it is difficult to give a definitive answer whether they appear in a movie, while for objective tags, such as "pirates" or "mafia", it is easier to give an answer.

Based on the results of the statistical analysis, we cannot make any strong claims regarding the difference in rating consistency between genre and other tag categories. However, based on the inconsistency of ratings (Figure 5), it seems that annotators mark the absence of genre tags consistently regardless of their abstractness (e.g. "fantasy" or "drama"). This indicates that it is possible that the genre tag category has lower inconsistency than subjective or misc categories, but we have not detected this difference, as we did not collect sufficient data for genre tags.

6.3 Tag Genome

The idea of Tag Genome is to represent an item as a vector, where each value corresponds to the degree to which a tag applies to

an item [38]. As according to our results (Figure 5), most ratings between 2 and 4 are inconsistent, it is possible that Tag Genome should instead focus on binary tags.

6.4 Item-tag rating inconsistency

We only investigated inconsistency of item-tag ratings over users. However, these ratings are probably also inconsistent over time and even within the same session, as is the case for item ratings [7, 36]. Similar to item ratings, item-tag ratings are also likely to be distributions of possible values [19]. It is possible that the way to reduce inconsistency is not to post process the data or collect more ratings, but to change the way ratings are collected [19]. Finally, similar to item ratings, inconsistency might not always be a negative property of item-tag ratings [19]. The inconsistency might be helpful in detecting item-tag pairs that are invalid, because annotators do not have consensus on them.

6.5 Tag subjectivity

Kotkov et al. [21] found that subjective tags are rated more inconsistently than objective ones and suggested that this difference could be due to different factors, such as annotators "forgetting certain parts of movies, using different scales or misunderstanding tags" [21], not only due to annotator opinions. Our findings seem to support this claim. Table 3 and 4 suggest that the main reasons for the difference between subjective and objective tags in terms of rating inconsistency are ease of rating and tag familiarity. Although subjective tags represent opinions of annotators and therefore vary among annotators, while objective tags represent facts and therefore should be the same for everyone [20, 21], however, this seems to have a low effect on rating inconsistency compared to other factors.

6.6 Limitations

This study has the following limitations: (1) a different and/or bigger sample of tags might produce different results; (2) our findings might only apply to popular movies and tags, as this was our selection criteria; (3) we selected factors based on interviews, but there might be other factors that affect rating inconsistency as well.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we investigated movie-tag rating inconsistency and the factors that can influence it. We conducted semi-structured interviews with annotators to enumerate reasons for rating inconsistency. We used these reasons to design a survey where we asked annotators to rate movie-tag pairs and provide additional information on their ratings. We ran the survey in Amazon Mechanical Turk and collected 6,070 ratings from 665 annotators.

We analyzed the collected dataset and found that ~45% of ratings were inconsistent. We also found that the easier annotators found it to rate a movie-tag pair, the less likely they were to provide inconsistent ratings. Furthermore, we found that subjective tags had higher inconsistency than objective ones, which was associated with the fact that annotators found objective tags more familiar and objective movie-tag pairs easier to rate.

In the future, we plan to extend this work in the following directions:

(1) We used tag definitions provided by annotators to exclude unreliable annotators (Section 4). An association analysis between tags and tag definitions might provide additional insights into why ratings are inconsistent and why there are differences between tag categories.
(2) We can detect movie-tag pairs to which annotators are likely to give inconsistent ratings. For this purpose, we can use ratings alone or in combination with tag or movie features extracted in [21, 38].
(3) We can further investigate whether rating inconsistency is mostly caused by tags, movies or their combinations by using different models.

ACKNOWLEDGMENTS

We would like to thank organizations that supported this work: the Academy of Finland, grant 309495 (the LibDat project) and the Academy of Finland Flagship programme: Finnish Center for Artificial Intelligence FCAI.

REFERENCES

[1] [n. d.]. Amazon Mechanical Turk. mturk.com. [Online; accessed 13-October-2021].
[2] [n. d.]. Internet Movie Database. https://imdb.com/. [Online; accessed 13-October-2021].
[3] [n. d.]. MovieLens. Non-commercial, personalized movie recommendations. https://movielens.org/. [Online; accessed 13-October-2021].
[4] [n. d.]. Netflix. https://www.netflix.com/. [Online; accessed 13-October-2021].
[5] [n. d.]. Quora. https://www.quora.com/. [Online; accessed 13-October-2021].
[6] [n. d.]. YouTube. https://www.youtube.com/. [Online; accessed 13-October-2021].
[7] Xavier Amatriain, Josep M Pujol, and Nuria Oliver. 2009. I like it... i like it not: Evaluating user ratings noise in recommender systems. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 247–258.
[8] Xavier Amatriain, Josep M Pujol, Nava Tintarev, and Nuria Oliver. 2009. Rate it again: increasing recommendation accuracy by user re-rating. In *Proceedings of the third ACM conference on Recommender systems*. 173–180.
[9] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction* 12, 4 (2002), 331–370.
[10] Cornelia Caragea, Florin Bulgarov, Andreea Godea, and Sujatha Das Gollapalli. 2014. Citation-enhanced keyphrase extraction from research papers: A supervised approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1435–1446.
[11] Michael Castelluccio. 2006. The music genome project. *Strategic Finance* (2006), 57–59.
[12] Joaquin Derrac and Steven Schockaert. 2015. Inducing semantic relations from conceptual spaces: a data-driven approach to plausible reasoning. *Artificial Intelligence* 228 (2015), 66–94.
[13] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382.
[14] Manish Gupta, Rui Li, Zhijun Yin, and Jiawei Han. 2010. Survey on social tagging techniques. *ACM Sigkdd Explorations Newsletter* 12, 1 (2010), 58–72.
[15] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (2015), 19:1–19:19.
[16] DP He, ZL He, and C Liu. 2020. Recommendation Algorithm Combining Tag Data and Naive Bayes Classification. In *2020 3rd International Conference on Electron Device and Mechanical Engineering (ICEDME)*. IEEE, 662–666.
[17] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Trans. Inf. Syst.* 22, 1 (Jan. 2004), 5–53. https://doi.org/10.1145/963770.963772
[18] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas. 1995. Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 194–201.
[19] Kevin Jasberg and Sergej Sizov. 2020. Human uncertainty in explicit user feedback and its impact on the comparative evaluations of accurate prediction and personalisation. *Behaviour & Information Technology* 39, 5 (2020), 544–577.
[20] Jie Kang, Kyle Condiff, Shuo Chang, Joseph A Konstan, Loren Terveen, and F Maxwell Harper. 2017. Understanding how people use natural language to ask for recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 229–237.

[21] Denis Kotkov, Alexandr Maslov, and Mats Neovius. 2021. Revisiting the Tag Relevance Prediction Problem. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1768–1772. https://doi.org/10.1145/3404835.3463019

[22] Denis Kotkov, Alan Medlar, Alexandr Maslov, Umesh Raj Satyal, Mats Neovius, and Dorota Glowacka. 2022. The Tag Genome Dataset for Books. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22)*. 5. https://doi.org/10.1145/3498366.3505833

[23] Denis Kotkov, Jari Veijalainen, and Shuaiqiang Wang. 2020. How does serendipity affect diversity in recommender systems? A serendipity-oriented greedy algorithm. *Computing* 102, 2 (2020), 393–411.

[24] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 379–390.

[25] Kai Lei, Qiuai Fu, Min Yang, and Yuzhi Liang. 2020. Tag recommendation by text classification with attention-based capsule network. *Neurocomputing* 391 (2020), 65–73.

[26] Benedikt Loepp, Tim Donkers, Timm Kleemann, and Jürgen Ziegler. 2019. Interactive recommending with tag-enhanced matrix factorization (TagMF). *International Journal of Human-Computer Studies* 121 (2019), 21–41.

[27] Suman Kalyan Maity, Abhishek Panigrahi, Sayan Ghosh, Arundhati Banerjee, Pawan Goyal, and Animesh Mukherjee. 2019. DeepTagRec: A content-cum-user based tag recommendation framework for stack overflow. In *European Conference on Information Retrieval*. Springer, 125–131.

[28] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[29] Hanh TH Nguyen, Martin Wistuba, Josif Grabocka, Lucas Rego Drumond, and Lars Schmidt-Thieme. 2017. Personalized deep learning for tag recommendation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 186–197.

[30] Tien T Nguyen, Daniel Kluver, Ting-Yu Wang, Pik-Mai Hui, Michael D Ekstrand, Martijn C Willemsen, and John Riedl. 2013. Rating support interfaces to improve user experience and recommender accuracy. In *Proceedings of the 7th ACM Conference on Recommender Systems*. 149–156.

[31] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web*. Springer, 325–341.

[32] Nirmala Pudota, Antonina Dattolo, Andrea Baruzzo, Felice Ferrara, and Carlo Tasso. 2010. Automatic keyphrase extraction and ontology mining for content-based tag recommendation. *International Journal of Intelligent Systems* 25, 12 (2010), 1158–1186.

[33] Gollam Rabby, Saiful Azad, Mufti Mahmud, Kamal Z Zamli, and Mohammed Mostafizur Rahman. 2020. Teket: a tree-based unsupervised keyphrase extraction technique. *Cognitive Computation* 12, 4 (2020), 811–833.

[34] Alan Said and Alejandro Bellogín. 2018. Coherence and inconsistencies in rating behavior: estimating the magic barrier of recommender systems. *User Modeling and User-Adapted Interaction* 28, 2 (2018), 97–125.

[35] Alan Said, Brijnesh J Jain, Sascha Narr, and Till Plumbaum. 2012. Users and noise: The magic barrier of recommender systems. In *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 237–248.

[36] Sergej Sizov. 2016. Assessment of rating prediction techniques under response uncertainty. In *Proceedings of the 8th ACM Conference on Web Science*. 363–364.

[37] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[38] Jesse Vig, Shilad Sen, and John Riedl. 2012. The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. *ACM Trans. Interact. Intell. Syst.* 2, 3, Article 13 (Sept. 2012), 44 pages. https://doi.org/10.1145/2362394.2362395

Table 5: Appendix. Selected tags and their tag categories.

Objective	Subjective	Genre	Misc
aliens	artistic	action	culture clash
animation	atmospheric	adventure	cyberpunk
christmas	bittersweet	art house	dystopia
comic book	boring	comedy	end of the world
drugs	disturbing	documentary	environmental
holocaust	funny	drama	existentialism
india	great soundtrack	fantasy	futuristic
mafia	hilarious	film noir	hallucinatory
martial arts	humorous	historical	insanity
nonlinear	inspirational	horror	justice
pirates	original	mockumentary	philosophical
prison	overrated	musical	plot twist
religion	predictable	mystery	psychological
robots	quirky	romance	self discovery
serial killer	stupid	satire	social commentary
space	stylized	sci-fi	steampunk
superhero	surreal	thriller	suspense
time travel	tense	western	twist ending
violence	thought-provoking		twists & turns
war	unique		
world war ii	visually appealing		
	witty		