Expectation vs reality: a qualitative study of movies expectation and its impact on reception

<This paper is a draft: it contains only the design of the research>

Eric Camellini
Delft University of Technology
Delft, The Netherlands
eric.camellini@gmail.com

ABSTRACT

Building effective movie recommender systems and box office prediction models is a difficult task, mainly because of the biases that affect the data that they use. In this work I define the *movie expectation bias*, an effect that could affect the quality of these models; it can be defined as the bias introduced by the difference between the way in which users evaluate a movie and their expectation before seeing it.

The aim of this study is to gain insight on this effect and to inspire further work that could better quantify it and model it. To do that I propose a technique to spot this effect using sentiment analysis on microblogging messages, and I apply this method on different recent movies to see if this bias exists or not. To understand if the effect modeled is really the one that I defined I also gather users opinions through a poll and check if they confirm the results obtained or not (methodological triangulation).

Results show that ... <RESULTS PLACEHOLDER>.

CCS Concepts

•Information systems → Web and social media search; *Web searching and information discovery;*

Keywords

Movie; Bias; Recommender system; Rating; Box Office; Prediction

1. INTRODUCTION

Movie industry has a great impact on the US economy: it contributes $40 billions to over 330,000 businesses annually, along with close to $16 billion to federal and state tax coffers [1]. Services like Netflix[1], that provide on-demand movies and TV series, are becoming more and more popular and are now part of everyday life. For these reasons a lot of past studies focused on *movie box office prediction* [12, 11, 5, 27, 17] and on improving *movie recommender systems* for online and on-demand services [13, 20, 26].

Most of the box office predictors are based on historical data, movie reviews and investment in advertising [27, 12, 17], while classic recommender systems base their models on users past behaviour and content similarity between different movies (e.g. Collaborative filtering or Content Based techniques) [13, 20, 26].

Unfortunately, these are difficult tasks; there are many techniques available, all with different limitations [3, 8], and many evaluation approaches that can lead to contrasting results depending on the application [16] (e.g. rating prediction, ranking, etc.). One of the main difficulties in this field comes from *biases* related to users behaviour (e.g. optimism or pessimism in rating movies) and to movies characteristics (e.g. popularity, critical reception and reviews, etc.).

Spotting and modeling these biases is an open problem; each one of them is often addressed and solved in a different, application dependant, way. Modern approaches in both fields (i.e. prediction and recommender systems) use data gathered from the web and the social media (e.g. hybrid recommenders [8]) to better capture the general opinion about a movie, and its critical reception [5, 11, 10, 21]. For example, the critics and users reviews impact on the movie reception (a known bias) can be included in the model by gathering data from on-line services such as IMDb[2] or Metacritic[3] [9].

In this study I define a bias that can affect both box office predictors and movie recommender systems: the *expectation bias*. This bias can be intuitively described as the "expectation vs reality" [4] effect applied to movie reception. If people have a high expectation for a movie and after its release this expectation is not met, there can be a negative impact on the reviews and on the users ratings, meaning that the movie could be rated lower than it deserves. In the same way, the opposite phenomenon holds. Furthermore, I propose a method to spot and model this effect: (1) gather Tweets from Twitter[5] before and after the movie release; (2) perform sentiment analysis on them; (3) analyse the average sentiment trend to see if there is a significant change.

The goal of this study is to gain understanding of the

[1]www.netflix.com/

[2]www.imdb.com

[3]www.metacritic.com

[4]a famous Internet meme: knowyourmeme.com/memes/expectation-vs-reality

[5]www.twitter.com

ACM ISBN .

DOI:

expectation bias and to determine if it can be modeled using sentiment analysis: to do so I apply the described technique on a set of recent and popular movies, and I verify if the results obtained show a real effect triangulating them with the users explicit opinion, gathered through a Twitter poll.

Results show that ... <RESULTS PLACEHOLDER>.

2. RELATED WORK

Recommender systems are an active research field, and have been so for the last decade. The research in this field started in the late nineties [25], and movies recommendation was one of the first focuses [13]; despite that fact, movie-related studies started growing exponentially only after Netflix launched the Netflix Prize [6] in 2006. The outcomes of this competition between researchers showed how challenging is to build recommender systems [7], especially because it is a task that highly depends on the items to be recommended and on the effects that item and user biases have on the data.

One of the first effective solutions to the prize was proposed by Bell et al. [6]: in a nutshell, they improved the classic collaborative filtering item-based technique [26] by including in the model the basic user and item biases (i.e. user trend to give good or bad ratings, item popularity bias, etc.), that were called *global effects*.

All the effective recommender systems nowadays take some biases into account [3, 16], and many studies focus only on bias detection and modeling [19, 23, 31, 2]. Biases have also shown to be important for box office prediction systems: past studies showed that it is effective to take into account the popularity of the movie and the effects of the reviews on the users opinion [15, 9, 24]. To my knowledge, none of the previous studies focused on the movie expectation bias, as defined in this work.

As a tool for this research we decided to use sentiment analysis; the usage of this technique is not new to recommender systems [4, 18]. In particular, Yadal et al. [30] used sentiment analysis to address a prejudice bias in movie reviews, so this work is not the first one that tries to link users sentiment with bias modeling.

The data that we use is gathered from the Twitter microblogging service; social media are in general widely use in modern studies. In particular, data from Twitter was previously used both to build recommender systems [22, 10, 21] and to perform box office prediction [11, 5], even though previous research shows that this practice has also many limitations [29].

3. METHODOLOGY

The main goal of this work is to gain insight into the problem of determining whether a given movie met the expectation of the people or not. The reason for doing that is that this difference between expectation and reality could bias the way in which people evaluate this movie, therefore the effectiveness of movie recommender systems and box office prediction models; I define this difference as the *movie expectation bias*.

My investigation can be defined through the following research question:

RQ : Can tweet sentiment analysis be used to understand whether a movie met the expectation of the people or not?

3.1 Proposed approach

To reach the goal of this research, I propose a methodology that can be used to model the movie expectation bias using sentiment analysis on data from the Twitter microblogging platform, and I analyse the results of applying this technique to different movies (see Section 4.1). The intuition behind this approach is that the expectation of a movie and the real opinion about it can be linked to the sentiment (positive or negative) of the people when they speak about it before and after seeing it. In a nutshell, for a given movie M I perform the following steps:

1. crawl the tweets for one week before the release of M and for one week after it (see Section 3.2);
2. perform sentiment analysis to classify the gathered tweets as positive or negative (see Section 3.3);
3. evaluate if the percentages of positive and negative tweets before and after the movie release are significantly different, and how do they differ (see Section 3.4).

The following Sections describe each one of these steps in a more detailed way.

3.2 Tweet Crawling and filtering

The first step of this research is crawling and filtering the Tweets that speak about the given movie M, in order to build the dataset for the analysis. To gather the Tweets I decided to use the Twitter Search API [7]: it is useful to find the messages that speak about a given topic and it already filters them since the it is focused on relevance and not on completeness. The extracted tweets are then used to build the datasets for the movie M: a row of the dataset contains the tweet itself together with its date; this makes possible to filter the ones published before and after the movie release in an easy way.

3.3 Sentiment Analysis

To perform sentiment analysis on the tweets I decided to use the Sentiment140 [8] API, built around the technique described by Go et al. [14]. With this API is possible to classify a given tweet as positive or negative. For a given movie, this sentiment information is added as a new column to the dataset that contains the related tweets (described in Section 3.2) and the resulting new dataset is what I use for the evaluation (Section 4).

3.4 Expectation and reception

To determine if a movie met the expectation of the people I decided to look at the general sentiment of the Tweets with respect to it, before and after its release. I assume that if a microblogging message speaks in a positive way about a movie before its release, then the author has a high expectation of it. In the same way, I assume that if a message speaks in a positive way about a movie after its release, then the author opinion about the movie is good. These are strong assumptions for many reasons (described in Section 4.5) and are obviously not always true, but since I am only interested

[6]http://www.netflixprize.com/

[7]dev.twitter.com/rest/public/search

[8]http://help.sentiment140.com/api

in capturing the general trend and in gaining some insights about the expectation bias, they can be considered valid for this work.

Based on these assumptions I define:

Expectation of the movie M: *proportion of positive tweets about M posted during the week that precedes its release;*

Reception of the movie M: *proportion of positive tweets about M posted during the week that follows its release;*

For example, if the 60% of the tweets collected for the movie M during the week before its release are positive, then the expectation of M is 0.6, meaning that the 60% of the related tweets speak positively about M before seeing it. In the same way, if these tweets were posted during the week after its release, then the reception of M is 0.6. Comparing expectation and reception for a movie should then give an indication of the general sentiment trend and on how it changes after its release. My interpretation of the relation between these two metrics is the following:

- if expectation and reception don't differ significantly then the movie generally met the expectation;
- if expectation is significantly higher than reception then the movie is considered generally worse than expected;
- if expectation is significantly lower than reception then the movie is considered generally better than expected;

4. EVALUATION

4.1 Study subjects and dataset

For this study I selected to use the five most popular movies (accordingly to IMDb) that were released between 15/12/2015 and 15/01/16. For each one of them I extracted the tweets as described in Section 3.2 and I created the dataset to be used for the evaluation by adding the sentiment column, as described in Section 3.3. The details of the resulting datasets can be seen in Table 1.

Movie	Tweets before release	Tweets after release
...	...	...

Table 1: Movie datasets sizes.

4.2 Experiment 1: Movie expectation and reception

4.2.1 Experiment design

In a nutshell, to determine if a movie met the expectations or not I applied the solution described in Section 3.4 on a sample of the dataset obtained after the sentiment analysis step, described in Section 3.3, using a sample size of 1000. The detailed procedure is the following:

1. sample 1000 tweets from the ones before the release of the movie and use them to compute the *expectation*;
2. sample 1000 tweets from the ones after the release of the movie and use them to compute the *reception*;
3. use t-test to verify if expectation and reception differ significantly, with a 0.05 significance level;

4.2.2 Results

Movie	Expectation	Reception	p-value
...	...	...	...

Table 2: Expectation and reception, with result of the significance test.

As shown in Table 2, ... <RESULTS PLACEHOLDER>.

4.3 Experiment 2: Methodological triangulation

4.3.1 Experiment design

In this second experiment I performed a *methodological triangulation* to qualitatively verify the results of the first part of the study and to see if the modeled effect is real or not. To do that I created a Twitter Poll for every selected movie using the dedicated functionality (introduced late 2015). The poll for a given movie M asked to the users the following question:

"Did M met your expectation?"

And allowed the users to select one of the following answers:

1 "Yes, and it was better than expected";

2 "Yes, it did";

3 "No, it didn't".

The data gathered through this poll can then be used to check if the users explicit opinion reflects the results of the first experiment.

4.3.2 Results

	Answer		
Movie	1	2	3
...	...	...	...

Table 3: Poll outcome.

Table 3 shows that ... <RESULTS PLACEHOLDER>.

4.4 Discussion

From the results of the two experiments it is possible to see that ... <RESULTS PLACEHOLDER>.

4.5 Threats to Validity

The described methodology relies on the Twitter Search and the Sentiment140 APIs, thus it inherits all the threats to the validity of the information that they provide are also threats to the validity of this research. For what concerns the first API, the main threat is related to the Tweet relevance: it could erroneously filter out tweets that are relevant for this study and vice versa. A solution to that problem could be to use the Twitter Streaming API[9] instead and to filter the tweets using a relevance-based technique like the one described by Tao et al [28]. The threat related to the Sentiment API is the fact that it could always provide wrong results, as all the classification-based techniques built to work on unstructured data.

[9]dev.twitter.com/streaming/overview

Many threats to the validity of this experiment come also from the collected tweets themselves, from the technique that I developed and from the definitions and assumptions that I made. For example, not all the Tweets that speak about a movie after its release are posted by people that already saw it; furthermore, these tweets are often affected by other biases that come from the first movie reviews and from other tweets or sources of information and this could also affect the general expectation day by day.

Another effect that could have threaten the results of this study is the fact that our measurements of reception and expectation (as defined in Section 3.4) could not differ from the proportion of positive sentiment in randomly selected Tweets, and this should be taken into account. Some threats also derive from the poll: in particular, the fact that the subjects that answered are different from the ones that tweeted the gathered messages.

Eventually, I don't claim these results to be generalizable, since they reflect part of the behaviour of two small portions of the population (i.e. people that are on twitter and tweeted about a specific movie or answered to the corresponding poll).

5. CONCLUSION & FUTURE WORK

In this work I defined a new bias that can affect movie recommender systems and box office predictors: the *movie expectation bias*. I also showed a technique that could be used to model it for a specific movie based on analysing the sentiment trend in the tweets that speak about it.

The aim of this work was to gain insight on this effect in a qualitative way, and to do that I applied the described technique on different movies and triangulated its outcomes with real users opinions gathered using a poll, to see if they reflect the obtained results and if the effect modeled is real or not.

The results of these experiments show that ... <RESULTS PLACEHOLDER>.

I think that this study cannot yet be used to include this effect in a real recommender system or prediction model, but I contributed to the past research in this field by defining a new bias and by proposing how to spot and model it.

Future work should improve the tweet filtering step, trying to distinguish which tweets come from people that already saw the movie after its release, and should define a sentiment analysis technique specific for movie tweets, instead of using an API: this because movie tweets use a specific language and set of keywords.

Furthermore, the described technique should be tested on more movies and in a quantitative way: for example, this effect could be included in a prediction model to see if this results in a significant improvement or not. The technique should also be expanded in order to include biases that can affect it, such as critic reviews, general change in the expectation after the release, movie popularity etc.

6. REFERENCES

[1] Motion picture association of america - creating jobs. http://www.mpaa.org/creating-jobs/. Accessed: September 15, 2020.

[2] G. Adomavicius, J. Bockstedt, S. Curley, and J. Zhang. De-biasing user preference ratings in recommender systems. In *RecSys 2014 Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2014)(Foster City, CA, USA, 2014)*, pages 2–9, 2014.

[3] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[4] D. H. Alahmadi and X.-J. Zeng. Ists: Implicit social trust and sentiment based approach to recommender systems. *Expert Systems with Applications*, 42(22):8840–8849, 2015.

[5] S. Asur, B. Huberman, et al. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[6] R. M. Bell and Y. Koren. Improved neighborhood-based collaborative filtering. In *KDD-Cup and Workshop*, pages 7–14. ACM press, 2007.

[7] R. M. Bell and Y. Koren. Lessons from the netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007.

[8] R. Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[9] P. K. Chintagunta, S. Gopinath, and S. Venkataraman. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Science*, 29(5):944–957, 2010.

[10] S. Dooms, T. De Pessemier, and L. Martens. Movietweetings: a movie rating dataset collected from twitter. In *Workshop on Crowdsourcing and human computation for recommender systems, CrowdRec at RecSys*, volume 2013, page 43, 2013.

[11] J. Du, H. Xu, and X. Huang. Box office prediction based on microblog. *Expert Systems with Applications*, 41(4):1680–1689, 2014.

[12] J. Eliashberg, J.-J. Jonker, M. S. Sawhney, and B. Wierenga. Moviemod: An implementable decision-support system for prerelease market evaluation of motion pictures. *Marketing Science*, 19(3):226–243, 2000.

[13] S. Ghosh, M. Mundhe, K. Hernandez, and S. Sen. Voting for movies: the anatomy of a recommender system. In *Proceedings of the third annual conference on Autonomous Agents*, pages 434–435. ACM, 1999.

[14] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12, 2009.

[15] T. S. Gruca and J. E. Berg. Public information bias and prediction market accuracy. *The Journal of Prediction Markets*, 1(3):219–231, 2012.

[16] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[17] M. Joshi, D. Das, K. Gimpel, and N. A. Smith. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The*

2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 293–296. Association for Computational Linguistics, 2010.

[18] A. Koukourikos, G. Stoitsis, and P. Karampiperis. Sentiment analysis: a tool for rating attribution to content in recommender systems. In *Proceedings of the 2nd Workshop on Recommender Systems in Technology Enhanced Learning*, pages 61–70, 2012.

[19] S. Krishnan, J. Patel, M. J. Franklin, and K. Goldberg. A methodology for learning, analyzing, and mitigating social influence bias in recommender systems. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 137–144. ACM, 2014.

[20] B. N. Miller, I. Albert, S. K. Lam, J. A. Konstan, and J. Riedl. Movielens unplugged: experiences with an occasionally connected recommender system. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 263–266. ACM, 2003.

[21] A. Oghina, M. Breuss, M. Tsagkias, and M. de Rijke. Predicting imdb movie ratings using social media. In *Advances in information retrieval*, pages 503–507. Springer, 2012.

[22] E. Otsuka, S. A. Wallace, and D. Chiu. Design and evaluation of a twitter hashtag recommendation system. In *Proceedings of the 18th International Database Engineering & Applications Symposium*, pages 330–333. ACM, 2014.

[23] S. Piramuthu, G. Kapoor, W. Zhou, and S. Mauw. Input online review data and related bias in recommender systems. *Decision support systems*, 53(3):418–424, 2012.

[24] D. A. Reinstein and C. M. Snyder. The influence of expert reviews on consumer demand for experience goods: A case study of movie critics*. *The journal of industrial economics*, 53(1):27–51, 2005.

[25] P. Resnick and H. R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[26] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

[27] S. M. Shugan and J. Swait. Enabling movie design and cumulative box office predictions using historical data and consumer intent-to-view. *University of Florida, working paper*, 2000.

[28] K. Tao, F. Abel, C. Hauff, and G.-J. Houben. What makes a tweet relevant for a topic. *Making Sense of Microposts (# MSM2012)*, pages 49–56, 2012.

[29] F. M. F. Wong, S. Sen, and M. Chiang. Why watching movie tweets won't tell the whole story? In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, pages 61–66. ACM, 2012.

[30] K. Yadav, S. K. Yadav, and S. Gupta. Detecting sentiments from movie reviews by integrating reviewers own prejudice. *Global Journal of Computer Science and Technology*, 15(2), 2015.

[31] X. Zhao, Z. Niu, and W. Chen. Opinion-based collaborative filtering to solve popularity bias in recommender systems. In *Database and Expert Systems Applications*, pages 426–433. Springer, 2013.