Box Office Sales and Social Media: A Cross-Platform Comparison of Predictive Ability and Mechanisms

Matthias Bogaert[a], Michel Ballings[b], Dirk Van den Poel[a], Asil Oztekin[c]

[a] *Ghent University, Department of Marketing, Innovation, and Organisation,Tweekerkenstraat 2, 9000 Ghent, Belgium*
[b] *The University of Tennessee, Department of Business Analytics and Statistics, 916 Volunteer Blvd., 249 Stokely Management Center, 37996 Knoxville, TN, USA*
[c] *The University of Massachusetts Lowell, Manning School of Business, Department of Operations and Informations Systems, 72 University Avenue, Pulichino Tong Building, 01854 Lowell, MA, USA*

Abstract

This paper aims to determine the power of social media data (Facebook and Twitter) in predicting box office sales, which platforms, data types and variables are the most important and why. To do so, we compare several models based on movie data, Facebook data, and Twitter data. We benchmark these model comparisons using various prediction algorithms. Next, we apply information-fusion sensitivity analysis to evaluate which variables are driving the predictive performance. Our analysis shows that social media data significantly increases the predictive power of traditional box office prediction models. Facebook data clearly outperform Twitter data and including user-generated content next to marketer-generated always improves predictive power. Our sensitivity analysis reveals that volume and valence based combination variables pertaining to Facebook comments are the most important variables. Furthermore, we provide an in-depth analysis of the potential mechanisms driving differential predictive ability of Facebook and Twitter. Our findings suggest that Twitter has less of an impact on box office sales than Facebook because Twitter users have less source credibility than Facebook users. Our results are important for practitioners, marketers and academics who want to employ social media data for box office sales predictions.

Keywords: Predictive analytics, social media, machine learning, data mining, box office sales

1. Introduction

Movies are complex products [18] and because of this complexity the movie industry involves a lot of risk. To mitigate these risks, decision makers such as producers, investors, distributors, exhibitors and marketers [20] try to predict box office revenues to optimize production and marketing [68]. Decision makers need predictive models in the pre-production phase to help them make decisions pertaining to production factors such as budget [29], and casting. Correspondingly, in the post-production, pre-release phase, decision makers need predictive models to optimize marketing spend and advertising budgets [68]. In this paper we focus on the latter.

Email addresses: Matthias.Bogaert@kuleuven.be (Corresponding author) (Matthias Bogaert), Michel.Ballings@utk.edu (Michel Ballings), Dirk.VandenPoel@UGent.be (Dirk Van den Poel), Asil_Oztekin@uml.edu (Asil Oztekin)

Marketing and advertising are of great importance in the movie industry. Movie marketing accounts for approximately 30% of the total production cost, on average $35.9 million per movie in 2007 [31]. Research shows that marketing can save lower-quality movies from failure [42]. Similarly, advertisement budgets may be reduced to achieve a given market-performance level if high box office revenues are expected [85]. Given the high financial stakes involved in marketing motion pictures [30], and hence the importance of these decisions, accurate box office predictions are desired. However, box office prediction has been shown to be a task of great difficulty [20] and understanding the predictive ability of new data sources is therefore important.

Since the rise of social media, substantial research has been conducted on the relationship between social media and movie sales [24]. Most of these studies found that user-generated content (UGC) such as online word-of-mouth (WOM), is one of the most important indicators of sales [16, 56]. Asur and Huberman [5] concluded in their study that online WOM has more predictive power than other, more traditional data sources such as the Hollywood Stock Exchange index, which is a virtual stock exchange for the entertainment industry. These findings provide important insights for practitioners since it allows them to focus on the most influential elements of online WOM to boost their revenues. For example, research regarding the influence of chatter on Twitter on movie sales has revealed that the number of tweets and positive tweets ratio are important influencers of box office sales [69].

While research concerning social media and box office sales has advanced to some extent, it still suffers from four main limitations. First, whereas the power of Twitter to predict movie sales has been studied extensively, less attention has been paid to Facebook. This is unfortunate, since Facebook contains a great number of potentially interesting predictors of movie watching behavior [8]. Therefore, it is important to know how both data sources perform in predicting box office sales. Second, previous research mainly focused on the impact of UGC on box office sales, while disregarding marketer-generated content (MGC). Nevertheless, research has shown that both UGC and MGC on brand page communities impact consumer purchase behavior [38]. Since both UGC generates considerable amounts of data compared to MGC, collecting and parsing UGC from both Facebook and Twitter can be intractable. Moreover, users of these predictive systems want a parsimonious model. Having to insert too many values in the model before it can generate the predictions would be considered too cumbersome. Third, previous studies have studied social media data in isolation of traditional movie characteristics data (MOV). However, to correctly evaluate the value of social media data models should control for movie characteristics [51]. To the best of our knowledge, no study has collectively included movie data, UGC, and MGC in their box office predictions and conducted a comprehensive analysis of Facebook and Twitter. Fourth, since current models are absent of social media data, or only analyze Twitter data and not Facebook data and use a narrower set of algorithms [3, 53, 47], no study has evaluated the relative importance UGC and MGC features originating from multiple social media data sources in predicting box office sales. Hence, this leaves several important questions unanswered: (1)‘What is the

predictive ability of social media data in estimating box office sales over and above traditional movie data?', (2) 'Which platform (Facebook or Twitter) is more predictive of box office sales and why?', and (3) 'Which variables are the most important predictors of box office sales?'.

This paper contributes to literature in several ways. First, we predict box office sales for 218 movies using traditional movie characteristics data (MOV) combined with Facebook and Twitter. To ensure that our results are robust, we benchmark these model comparisons using seven algorithms: regularized linear regression, k-nearest neighbors, decision trees, bagged trees, random forest, gradient boosting and neural networks. The second contribution entails exploring theoretical mechanisms driving differential predictive ability between Twitter and Facebook. In doing so we offer theoretical background, and confirm our hypothesis with additional analyses. These additional analyses use singular value decomposition to control for multicollinearity in our prediction models such that we can uncover the true relationship between predictor and response [70]. The third contribution entails applying information-fusion sensitivity analysis to evaluate which variables from which platform (i.e., Facebook or Twitter) from which data type (i.e., UGC, and MGC) are driving predictive performance [8]. To demonstrate these contributions, we introduce a social media analytical methodology, which is an enhancement of the CRISP-DM framework [15]. To the best of our knowledge, this study is the first to conduct such a comprehensive and robust comparison between Twitter and Facebook as a data source for box office predictions while controlling for movie characteristics. Moreover, we are the first to thoroughly investigate the descriptive and predictive power of both Facebook and Twitter in regard to box office revenues based on an extensive set of UGC and MGC variables on top of movie characteristics and offer theory and analysis to explore potential mechanisms underlying differences in predictive ability.

The remainder of this paper is organized as follows. First, we provide an overview of the existing literature. Second, we discuss the methodological framework, extracted data, variables, algorithms, and information-fusion sensitivity analysis. Third, we describe our results comparing the predictive ability of the two platforms. After having established the differential predictive ability, we explore potential mechanisms underlying these differences across platform and movie type. Next, we perform sensitivity analysis to uncover the driving forces of predictive performance. Finally, we conclude a discussion and elaborate on the avenues for future research.

2. Literature review

Research on box office predictions has mostly studied two types of variables: movie characteristics and UGC, such as WOM [47]. The former includes variables such as the cast of the movie, the content of the movie and the release time of the movie [51]. The latter consists in the influence of WOM volume and valence on movie sales [27]. Together with the rise of social media, research concerning box office revenues and WOM has shifted from more traditional web 2.0 sites (e.g., Yahoo!Movies and blogs) to social network sites (SNS), such as Twitter and Facebook. The reasons for this shift are

manifold. First, Facebook and Twitter have a large user base with respectively 2.45 billion [34] and 321 million monthly active users [71]. Second, Facebook and Twitter allow companies to create their own customized Facebook and Twitter pages on which they can post their own promotional content. Facebook even allows companies to target a certain audience (e.g., users that live in New York and like to watch movies) [33]. Third, both platforms contain a lot of user-created and company-created content that have proven to have a significant impact on movie sales [24]. For the aforementioned reason, we decide to focus our study on Facebook and Twitter[1].

Studies on social media data and box office revenues can be categorized according to several dimensions: (1) whether they use movie, Facebook or Twitter data[2], (2) whether they include MGC and UGC, (3) whether they compare the predictive performance of both platforms, (4) whether they allow nonlinear effects, and (5) the number of movies they predict (Table 1). Studies including Twitter use the tweets about a movie to forecast movie sales [5]. For example, Rui et al. [69] found, using a dynamic panel model, that tweets expressing the intention to watch a movie have the strongest effect on movie sales. Moreover, they also found that people with more followers have a higher impact on revenues. Studies including Facebook data use information on the movie page to estimate movie sales [62]. For example, Ding et al. [24] examined the impact of a Facebook movie page like on box office sales and found that a 1% increase in pre-release likes leads to a 0.2% increase in opening week box office sales. Marketer-generated content (MGC) contains the volume and the valence of Facebook posts or Tweets created by the page owners (i.e., digital marketers of the focal firm) to increase engagement [38]. For example, the total number of Facebook posts refers to volume, whereas the average sentiment of a firm's Facebook posts relates to valence. User-generated content (UGC) often refers to the volume as well as the valence of online WOM about a certain movie [28]. For example, Asur and Huberman [5] use both the rate of the tweets (i.e., volume) and two sentiment measures (i.e., polarity and subjectivity) to model box office revenues. Next, studies comparing social media platforms assess which data hold the most predictive power. For example, Oh et al. [62] found that Twitter lost all predictive power of movie sales when Facebook data were entered in the model. However, they include the volume of UGC (e.g., the total talk on Facebook and the total tweets on Twitter) but neglect to add MGC as well as the valence of the online chatter. Finally, studies that allow for nonlinear effects do not assume that the relationship between box office sales and social media predictors are linear. For example, Oh et al. [62] use robust OLS, thereby assuming a linear relationship between box office sales and social media data. In contrast, Kim et al. [47] use support vector regression to model box office sales, thereby allowing for nonlinear boundaries.

Table 1 summarizes the literature on movie sales and social media data. From Table 1, it is clear that no study has included MGC and UGC in combination with movie data to predict box office sales,

[1]In the remainder of this article, we simply refer to both Facebook and Twitter as 'social media'.

[2]We solely focus on papers including movie data in combination with social media data.

accounted for nonlinear effects, and conducted an analysis of the value Facebook and Twitter (i.e., the dominant platforms in the marketplace). Facebook and Twitter are becoming more and more important as a tool for building brand equity and increase consumer engagement [80]. Applications such as Facebook Ads even allow firms to target specific audiences with their brand posts [33]. Due to these platforms' popularity, thousands of tweets, comments and likes are created every second. If a firm wants to collect all the available Twitter and Facebook data, they have to use the Twitter or Facebook API [35, 77]. Both platforms have their own types of data and data limits, so collecting, parsing and preparing data from both platforms can be unmanageable in the long run. When firms want to predict the box office success of their movie, they want to get the most accurate predictions as efficiently as possible. Hence, they want to know whether including social media effectively increases the performance of box office prediction models. Once a firm knows social media data is effective, a second question is which what type of data to gather and which platform has the highest impact on predictive performance. UGC have proven to have a significant impact on box office sales predictions [24, 47]. However, several other studies have shown that both volume and valence of MGC have a significant influence on key performance metrics, such as brand equity, brand sales and profitability [80, 38, 49]. Hence, amongst the clutter of UGC and MGC marketers want to know which type of data to include on top of traditional movie data to increase performance. For example, do we need to post a lot of content ourselves or do we need to focus on generating buzz from the users?

Table 1: Overview of box office prediction literature including Twitter and/or Facebook

Study	Platform	MGC		UGC		Compare platforms	Nonlinear effects	N(Movies)
		Volume	Valence	Volume	Valence			
Asur and Huberman [5]	Twitter			✓	✓			24
Reddy et al. [65]	Twitter			✓	✓			1
Wong et al. [84]	Twitter				✓			34
Apala et al. [2]	Twitter				✓			35
El Assady et al. [28]	Twitter				✓			20
Guàrdia-Sebaoun et al. [40]	Twitter			✓	✓			32
Jain [45]	Twitter			✓	✓			30
Rui et al. [69]	Twitter			✓	✓			63
Arias et al. [3]	Twitter				✓		✓	50
Du et al. [26]	Weibo			✓	✓			24
Hennig-Thurau et al. [43]	Twitter			✓	✓			105
Liu et al. [53]	Weibo				✓	✓		57
Gaikar et al. [37]	Twitter			✓	✓			14
Kim et al. [47]	pulseK [1]			✓	✓		✓	212
Ding et al. [24]	Facebook	✓						64
Oh et al. [62]	Twitter, Facebook			✓		✓		106
Baek et al. [6]	Twitter among others [2]			✓				145
Houston et al. [44]	Movie, Twitter, Facebook	✓		✓				254
Our study	**Movie, Twitter, Facebook**	✔	✔	✔	✔	✔	✔	**218**

[1] PulseK aggregates data from several social network services and performs sentiment analysis. It was explicitly mentioned that Twitter is part of this data set. Usage of Facebook is not mentioned.

[2] Their study also includes Yahoo!Movies, YouTube and blog posts as social media channels. Since we are only interested in whether the study includes Facebook or Twitter, these channels are not mentioned.

To fill this gap in literature, we study the predictive power of Facebook and Twitter using several prediction models, while accounting for movie characteristics. Next to UGC, we also include MGC, such as the number (and valence) of posts and the number (and valence) of tweets generated by the firm itself. We only include data prior to the release of the movie. Part of extant literature uses only pre-release data [51] and part of the literature uses pre and post release data [24, 47]. Both parts have different goals. The latter studies want to forecast box office sales the next week by using all the available data before that week [47]. The main issue with these after-release-features is that when attempting to predict box office gross in a real-world setting, producers would not have information available after release. Anybody who is interested in making such predictions would want to do so before the movie is released given that the goal is post-production pre-release marketing optimization. Hence, the goal of this paper is prediction, and therefore the logical choice is to only include pre-release data. To do so, we introduce a social media analytical approach consisting of two stages. The first stage contains the data collection, the feature engineering, the model estimation and model comparison. To compare both platforms we create several models for each platform. Our baseline model only uses movie data. Next, we augment this model once with MGC and once with UGC from Twitter and Facebook. In addition, we also include models containing both Facebook and Twitter as well as movie data, UGC, and MGC. The reason is that MGC and especially UGC induces a large computational overhead, and should only be collected when it significantly improves predictive performance. Thus, in addition to analyzing the added value of Facebook and Twitter, we also assess the added value of UGC and MGC. A previous study by Oh et al. [62] concluded that Twitter followers had a significant positive influence on movie sales when studied in isolation. However, when Facebook likes were introduced into the model, Twitter followers became insignificant and Facebook likes turn out to be significant. They argue that Facebook is more consumer-centric and information-rich than Twitter, thereby weakening the effect of Twitter on box office sales. Another study of Lo [55] argues that movie watchers rely more on Facebook than specialized sites such as Yahoo!Movies. Moreover, they also conclude that, overall, Facebook is considered a more important social network site than Twitter. Hence, it is possible that Facebook would be more significant in predicting box office revenues than Twitter. To ensure that our results are reliable, we compare both platforms using several algorithms: regularized linear regression (LR), k-nearest neighbors (KN), decision trees (DT), bagged trees (BT), random forest (RF), gradient boosting (GB) and neural networks (NN). We included these algorithms since they have been shown to have superior performance in predicting box office sales and firm performance in general [72, 20].

The second stage summarizes the information from both platforms and prediction models with a technique called information-fusion sensitivity analysis [64]. Information-fusion combines the knowledge of all prediction models in an unbiased and balanced fashion and determines which variables are the driving force of predictive performance [63]. Hence, it can be seen as an advanced way of measuring variable importances, since it determines the impact of a variable across all prediction models. In

agreement with previous literature, we believe that WOM (or more in general UGC) would be of major importance in comparison to MGC. User-generated content, and more specifically pre-release consumer buzz, shows a degree of interest and anticipation towards a certain movie [44]. According to consumer engagement behavior theory, liking or following a movie reflects intrinsic motivation and involvement, and commenting and replying reflects socializing and participation in the community and hence leads to higher movie sales [62]. Therefore, more UGC leads to higher box office revenues.

Research investigating the relationship between UGC and movie sales has mainly focused on online WOM. WOM influences movie sales through two mechanisms: the awareness and persuasive effect [54]. The former states that people can only consider products of which the existence they are aware. As the volume of online chatter increases, the awareness will increase and the movie will become a part of the potential customer's consideration set [69]. The latter helps people create their attitude and opinions towards the product through the information they receive, which in turn affects their purchase decision [27]. As the valence (or sentiment) of the tweet becomes significantly more positive, the persuasive effect and the probability to buy a product becomes larger [75]. A lot of researchers have focused on the effects of both volume and valence, but not all studies reach the same conclusions. On the one hand, Liu [54] found that volume, measured as the total number of WOM interactions on Yahoo!Movies, was the most important influencer of movie sales. They did not find a significant relationship between valence and sales. Wong et al. [84] came to the same conclusions. On the other hand, Chintagunta et al. [16] found that valence, and not volume, was the most important variable. Rui et al. [69] found that both volume and valence had an effect on box office revenues. Hennig-Thurau et al. [43] thoroughly tested the effect of valence and concluded that negative WOM dominates positive WOM and has a negative influence on early adoption. The major reason for these discrepancies are the broad range of alternatives to come up with volume and valence. Moreover, most studies only include their own volume or valence measure neglecting to test the performance of their measure against the existing alternatives. For example, Asur and Huberman [5] used a simple positive and negative tweet ratio for valence, while Kim et al. [47] employs the total number of emotional, positive and negative SNS mentions.

Extant research concerning social media and box office sales has not investigated the relationship between MGC and box office sales. However, other studies have demonstrated that, in addition to UGC, MGC also has an influence on several firm performance metrics, such as brand equity and acquisition [80], customer spending, cross-buying and profitability [49]. Compared to UGC, Goh et al. [38] found that both volume and valence of MGC drive consumer purchases, however to a lesser extent than UGC. The reason is that MGC influences consumer behavior only through the persuasive effect, whereas UGC impacts consumer purchase through informativeness and persuasiveness. In general, several studies have shown that the effect of UGC is more significant than MGC in explaining firm performance [1, 39].

In conclusion, we are the first to conduct such an extensive analysis of Facebook and Twitter within the context of box office sales. On the methodological side, we contribute to literature by analyzing a large collection of movies, among studies analyzing social media data, with a wide range of algorithms. Moreover, we introduce a generic social media analytical framework in a subsequent section that can help researchers and practitioners replicate our methodological approach in other similar settings. On the theoretical side, we contribute to literature by assessing which content type dominates box office revenues while using social media data. In the next section, we discuss our framework, materials and methods. In the next section we introduce our theoretical framework explaining the mechanisms underlying differential prediction performance of Twitter and Facebook.

3. Theoretical framework

We now explore the theoretical foundations that are driving the differential predictive ability between Twitter and Facebook. Extant literature states that the volume of online user-generated content (UGC), also referred to as chatter or pre-release buzz [44], is related to future product sales. For example, Dhar and Chang [22] find that future music album sales are positively correlated with the volume of blog posts about an album, Baek et al. [6] find that the number of tweets related to a movie is related to box office revenue, and Goh et al. [38] find a positive relationship between UGC and apparel purchase expenditures. Malthouse et al. [57] provide evidence that UGC that produces engagement increases purchase behaviors. They use the elaboration likelihood model to explain their finding by saying that if a user elaborates (i.e., comments or replies), there first needs to be relevance to a personal goal and motivation. Therefore, comments or replies are leading indicators of purchases. Viswanathan et al. [79] find that the number of tweets by users is positively linked to subsequent TV show viewing behavior. In sum, users who produce positive UGC (the poster) are themselves more prone to buying, and stimulate others (the readers) to buy.

The findings in the aforementioned studies are all based on linear models. One notable study looks into the UGC-sales relationship by exploring nonlinear relationships. By doing so, Maslowska et al. [58] find, contrary to popular belief, that more positive online reviews do not always translate into higher sales. The probability of purchase increases with review ratings to about 4.2-4.5 out of 5, but then decreases, and explain this as consumers perceiving this as being too good to be true. Social media platforms such as Twitter and Facebook are environments in which deception can be beneficial for the deceiver. Consumers are aware of the fact that, a movie marketer might potentially be inclined to generate fake positive signals about the movie [52], for example by purchasing positive pre-release buzz from buzz farms.

According to Warranting Theory, source credibility is a major component in whether consumers believe whether UGC is not to be trusted. The fact that profiles are online, calls into question the extent to which they reflect offline reality or seem authentic and trustworthy [81]. The guiding principle

in Warranting Theory is that users seek to understand who can manipulate comments and replies online [78]. The more users believe that these pieces of information are manipulated by the entity that is to gain from it, say the movie marketer, the less credible people perceive the information to be [78]. In other words, Warranting Theory predicts that replies on Twitter, and comments on Facebook, about a specific movie are only influential if the recipients of these replies and comments (i.e., other potential consumers of the movie) believe that the creators of these messages are not controlled by the movie marketer. Any clues that indicate that the online commenters are affiliated with the movie marketer, would reduce the influence of these comments [19]. Greater potential of misrepresentation would result in viewers of these comments and replies to more likely be skeptical of the presented information. Kim et al. [46] confirm that source credibility is an important driver of purchase probabilities.

The question now arises as to which platform, Twitter or Facebook, has higher source credibility. In this context, Signaling Theory provides clues as to why user profiles on Facebook might be more reliable and trustworthy than user profiles on Twitter [50]. Signaling Theory addresses the basic question of what keeps signals reliable in environments where deception can be beneficial to the deceiver. Of importance is implicit and explicit verifiability of the identity behind user profiles. Explicit verification refers to online clues that can be exactly verified. For example, if a user is tagged in pictures of friends, one can explicitly verify the identity of that user. Implicit verification refers to clues that cannot be directly linked to the user itself. For example, if a user provides his/her school, this cannot be exactly verified if there is no pictorial proof. The network structures of Twitter and Facebook are fundamentally different in terms of implicit and explicit verification since their inception. Twitter is a micro-blogging site that allows its users to publish (Tweet) and reply to short posts [73]. The focus is on the content sharing and entertainment, and there is limited biographical information available about the user. By default profiles are public and users only have to be identified by their username [74]. Twitter users also tend to follow only a few famous profiles, leading to a lot of users with a small network size and a few user with [13]. While Facebook is also used for publishing content, the network was created for users to produce elaborate profiles, feature personal information such as interests, photos, and preferences, and connect with users [73]. It is therefore also harder to grow one's network on Facebook: one will not share personal information with strangers. By default user profiles are set to private and users can only see personal information after having established a trusted link [74]. Previous research has also shown that Facebook users only tend to add people as a friend that they have met in real-life [32]. In other words, it is much harder, and much more laborious to produce a fake consistent user profile and establish a fake friend network on Facebook, than it is on Twitter. This also explains why the average network size on Facebook remains stable [13]. Given the two inherently different network structures on Facebook and Twitter, there are more signals about the source are available, consistency and honesty of the source is easier to police and verify [50], and therefore it would follow that signals produced on Facebook should be conceived as more reliable than signals produced on Twitter.

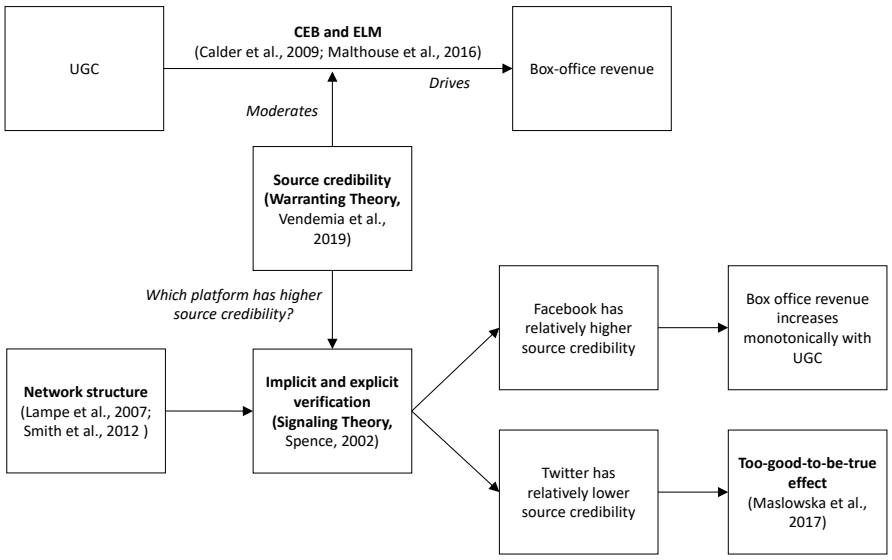


Figure 1: Schematic illustration of our theoretical framework

If the hypothesis that users on Facebook have higher source credibility than users on Twitter is true, then the impact on box office sales of user-generated content (UGC) on Facebook should be different than the impact of UGC on Twitter. More specifically, the prediction is that Twitter will be more likely to present a too-good-to-be-true effect as found in Maslowska et al. [58] (note that their application was on online reviews) when signals become more extreme. In the case of Facebook, there should be no too-good-to-be-true and box office revenues should increase with UGC. We summarize our theoretical work in Figure 1.

4. Methodology

4.1. Framework

The framework employed in our study is a holistic integration of the well-known CRISP-DM methodology [15]. The CRISP-DM framework is the most commonly used methodology in analytics and ensures robust results [63]. There are six steps in the process: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The main adaptation of our framework emerges in the data collection, data preparation and information-fusion sensitivity analysis. In our framework the four data sources (i.e., BoxOfficeMojo, OMDb, Facebook and Twitter) are crawled from the internet using the API.

The first step is the data collection. In this step the social media data are gathered from the Twitter and Facebook API. Movie data are collected from the OMDb API[3]. This is a restful API containing the same data as the International Movie Database (IMDb). Finally, the BoxOfficeMojo website is used to collect the gross box office revenues for the desired movies. The second step involves the inspection of the raw data sources (i.e., data understanding). The third step involves cleaning,

[3]www.omdbapi.com

handling and creation of the different basetables involving movie, Twitter and Facebook data sources. There are two different data preparation procedures. The first procedure involves numeric and time variables that do not require any text processing. The second procedure includes text and sentiment analysis. The output of this step is several basetables including movie, user-generated and marketer-generated content (separate or in combination). Next, for each basetable, 7 prediction models are built using 5 times two-fold cross-validation (5x2cv). Afterwards, the models are evaluated and compared against each other to determine the best platform and the best algorithm. Finally, information-fusion is applied to integrate the knowledge of all prediction models and movie, Facebook and Twitter variables. Using the fusion model, variable importances are assessed to uncover the driving forces of predictive performance.

4.2. Data

We extracted data from 218 movies released between January 2012 and December 2015 that had a verified[4] Facebook or Twitter page. The included movies are a mix of mainly big blockbusters (e.g., The Martian, The Revenant, Mad Max) and lesser known movies (e.g., Sisters, Goosebumps) that are released worldwide. We obtained data from the Facebook and Twitter pages from the start of their very existence until the release date of the movie. If we would allow information into the model from after the release of the movie, this would mean that we are violating the operational context, meaning we would not be able to draw conclusions about the predictive ability of Facebook and Twitter. For the same reason, we only selected movies until the end of 2015 because we wanted to be certain that the movies were out of theaters and thus reached their final gross box office revenues. To extract the information of the Facebook and Twitter pages we used the publicly available API [35, 77], to extract the movie data we used the OMDb API. The main advantage of using APIs is that the data is freely available to everybody. Note that with introduction of GDPR, extraction of Facebook data via the API is only possible for page administrators. However, movie producers often own several pages and therefore we still believe that they can replicate our results. Page owners can extract the same social media data as in our study from their dedicated APIs and use our social media analytical framework to implement their predictive models. Movie sales data (i.e., gross box office revenues) were collected via BoxOfficeMojo within the same time window [10].

The movie data consists of all the information that can be found in the International Movie Database (IMDB). For example, the release date, the main actors, directors, genre, and a plot synopsis of the movie are collected. MGC refers to all information on a Facebook page or Twitter wall posted by the page owners (in our case movie producers) [61]. For example, the average sentiment of posts and the total number of posts on the Facebook page refer to the valence and volume of MGC. User-generated

[4]Facebook and Twitter add a blue badge on verified pages. This means that Facebook or Twitter confirms that this is the authentic page for a movie.

Table 2: Descriptive statistics

Type of content		Total	Median	Min	Max
MGC					
	Facebook	76,671	248	47	1884
	Twitter	139,519	441	8	2992
UGC					
	Facebook	2,371,840	6155	127	122,999
	Twitter	25,156,997	33,622	815	3,439,413
Revenues ($)		-	23,621,057	25,480	356,461,711

content (UGC) consists of interaction initiatied by other users, like Facebook comments and Twitter replies and retweets [61]. For example, the average sentiment of replies or comments and the number of comments, replies or retweets are user-generated content. Table 2 presents descriptive statistics of the data sources. The 'Total' column gives the total amount of MGC or UGC collected on Facebook and Twitter across all 218 movies. For example, we collected 76,671 posts and 139,519 tweets in total across all movies and the median number of marketer-generated posts across all movies is 248 with a minimum of 47 posts and a maximum of 1884 posts for a movie. The final row of Table 2 shows the median and range of gross box office revenues. Since this distribution is skewed, we take the natural logarithm of the gross box office revenues as our dependent variable [69].

Based on these data types, we propose 5 data sets to compare both platforms (Table 3 first 5 rows). The baseline models include movie characteristics and MGC. The augmented models include MGC and UGC from Facebook and Twitter. This is motivated by the fact that collecting UGC induces a large computational overhead. On Facebook the API allows to collect the wall of a movie page. If one wants to collect the comments (UGC), one needs to sequence over all the collected wall posts and extract them individually. For Twitter the difference between MGC and UGC is even more significant, since not all the replies on the tweets should be gathered but also the retweets. For example, in Table 2 there are only 139,516 observations for MGC compared to 25,156,997 instances for UGC. In other words MGC can be collected without UGC, but UGC cannot be collected without collecting MGC. Therefore, it is of major importance for a company to know whether or not the collection of UGC is worth the effort. Next to these models we also added 2 models that augment the baseline model with both Facebook and Twitter data, one with UGC and one without UGC (Table 3 last 2 rows). These models aim to check whether the UGC has added value across both platforms. Table 3 summarizes the models used in this study including the data sources which they are composed of and the total number of variables. Note that Facebook and Twitter do not return the timestamps of the likes, shares, and favorites. Therefore, we only focus on the content of the posts (comments) and tweets (replies) since they contain timestamps.

Table 3: Overview models

Models	Movie	Facebook		Twitter		N(variables)
		MGC	UGC	MGC	UGC	
Movie	✓					33
Fb:base	✓	✓				55
Fb:plus	✓	✓	✓			77
Tw:base	✓			✓		55
Tw:plus	✓			✓	✓	97
FbTw:base	✓	✓		✓		77
FbTw:plus	✓	✓	✓	✓	✓	141

Note: Fb:base represents a model with movie and Facebook data with MGC, Fb:plus movie and Facebook data with MGC and UGC, Tw:base movie and Twitter data with MGC, Tw:plus movie and Twitter data with MGC and UGC, FbTw:base movie, Facebook and Twitter data with MGC, and FbTw:plus movie, Facebook and Twitter data with MGC and UGC.

4.3. Variables

Table 4 provides an overview of our movie and social media predictors together with the variable number, the definition and the data type and WOM type for our social media predictors. For our traditional movie features, we chose to follow the classification of Lash and Zhao [51] and include only pre-release predictors of box office sales. For example, the number of theaters cannot be exactly known before the release of the movie, hence this variable could contain information about our dependent and would violate our time window. Following the classification of Lash and Zhao [51] we include basic features about the 'what', the 'who', and the 'when'. 'Who' features tell something about who is involved in the movie. For example, we include whether a top actor or a top director is involved in the movie. 'What' features reflect meta-data about the movie and the plot. As meta-data, we included the 19 different genres (e.g., action, adventures) and the Motion Picture of America Association (MPAA) rating (e.g., R-rated, PG13) and the runtime of the movie. For the plot feature, we included the length of the plot synopsis. Finally, 'when' features are about the release date of the movie (e.g., the season).

For our social media predictors, we added all relevant variables as present in current literature as well as additional combinatorial variables. We note that we only included social media variables before the release of the movie. Hence, only variables that could be restricted in time, such as the number of comments and posts for Facebook and the number of tweets and replies and retweets for Twitter. This means that all variables are compared against the creation date of the object (e.g., a post, tweet, comment reply or retweet) in relation to the release date of the movie. By doing so our study focuses on pre-release social media indicators of box office sales. If we would allow variables after the release date, we would not be able to investigate whether social media really predicts box office sales.

The 'data' column in Table 4 refers to the data sources identified in Section 4.2, whereas the 'type' column describes whether the variable contains volume or valence information, or a combination of

both. For example, the number of positive posts before release is classified as a combination of volume and valence, whereas the percentage of positive posts before release and the average sentiment of the post is only valence.

Table 4: Overview of social media predictors

Nr.	Variable	Definition	Data	Type
	Movie			
1-19	Ind($Genre$)	Indicator of movie genre	-	-
20-26	Ind($Rating$)	Indicator of the MPAA film rating G	-	-
27	$Runtime$	The total runtime of the movie in minutes	-	-
28	Ind($TopActor$)	Indicator whether a top actor starred in the movie	-	-
29	Ind($TopDirector$)	Indicator whether the movie was directed by a top director	-	-
30-32	Ind($Season$)	Indicator in which season the movie was released	-	-
33	Length($plot$)	The number of characters present in the plot synopsis	-	-
	Facebook			
1	N($Posts$)	Number of posts before the release date	MGC	Vol
2	Pct($Posts$)	Percentage of the posts before the release date	MGC	Vol
3	N($Posts_{1week}$)	Number of posts posted from 1 week before the release date until the release date	MGC	Vol
4	Pct($Posts_{1week}$)	Percentage of posts posted from 1 week before the release date until the release date	MGC	Vol
5-7	N($SentPosts$)	Number of positive, neutral,negative posts before the release date	MGC	Comb
8	R($Posts$)	Ratio positive versus negative posts before the release date	MGC	Val
9-11	Pct($SentPosts$)	Percentage of posts before release that are positive, neutral, negative	MGC	Val
12-14	N($Sen, Posts_{1week}$)	Number of positive, neutral, negative posts from 1 week before the release date until the release date	MGC	Comb
15	R($Posts_{1week}$)	Ratio positive versus negative posts before release date	MGC	Val
16-18	Pct($SentPosts_{1week}$)	Percentage of positive, neutral, negative posts 1 week before release date until the release date	MGC	Val
19	Avg($SentPosts$)	Average sentiment of posts before the release date	MGC	Val
20	Avg($SentPosts_{1week}$)	Average sentiment of posts 1 week before the release date until the release date	MGC	Val
21	Avg($DailyPosts$)	Average number of posts per day before the release date	MGC	Vol
22	Avg($DailyPosts_{1week}$)	Average number of posts per day 1week before the release date until the release date	MGC	Vol
23	N($Comments$)	Number of comments before the release date	UGC	Vol

24	Pct($Comments$)	Percentage of comments before the release date	UGC	Vol
25	N($Comments_{1week}$)	Number of comments 1 week before the release date until the release date	UGC	Vol
26	Pct($Comments_{1week}$)	Percentage of comments 1 week before release until the release date	UGC	Vol
27-29	N($SentComments$)	Number of positive, neutral, negative comments before the release date	UGC	Comb
30	R($Comments$)	Ratio positive versus negative comments before the release date	UGC	Val
31-33	Pct($SentComments$)	Percentage of positive, neutral, negative comments before the release date	UGC	Val
34-36	N($SentComments_{1week}$)	Number of positive, neutral, negative comments 1 week before the release date until the release date	UGC	Comb
37	R($Comments_{1week}$)	Ratio positive versus negative comments 1 week before the release date until the release date	UGC	Val
38-40	Pct($PSentComments_{1week}$)	Percentage of positive comments 1 week before the release date until the release date	UGC	Val
41	Avg($SentComments$)	Average sentiment of comments before the release date	UGC	Val
42	Avg($SentComments_{1week}$)	Average sentiment of comments 1week before the release date until the release date	UGC	Val
43	Avg($DailyComments$)	Average number of daily comments before the release date	UGC	Vol
44	Avg($DailyComments_{1week}$)	Average number of daily comments 1 week before the release date until the release date	UGC	Vol
	Twitter			
1	N($Tweets$)	Number of tweets before the release date	MGC	Vol
2	Pct($Tweets$)	Percentage of the tweets before the release date	MGC	Vol
3	N($Tweets_{1week}$)	Number of tweets from 1 week before the release date until the release date	MGC	Vol
4	Pct($Tweets_{1week}$)	Percentage of tweets posted from 1 week before the release date until the release date	MGC	Vol
5-7	N($SentTweets$)	Number of positive, neutral negative tweets before the release date	MGC	Comb
8	R($Tweets$)	Ratio positive versus negative tweets before the release date	MGC	Val
9-11	Pct($SentTweets$)	Percentage of tweets before release that are positive, neutral, ,negative	MGC	Val
12-14	N($SentTweets_{1week}$)	Number of positive, neutral, negative tweets from 1 week before the release date until the release date	MGC	Comb
15	Ratio($Tweets_{1week}$)	Ratio positive versus negative tweets before release date	MGC	Val

16-18	Pct($SentTweets_{1week}$)	Percentage of positive, neutral, negative tweets 1 week before release date until the release date	MGC	Val
19	Avg($SentTweets$)	Average sentiment of tweets before the release date	MGC	Val
20	Avg($SentTweets_{1week}$)	Average sentiment of posts 1 week before the release date until the release date	MGC	Val
21	Avg($DailyTweets$)	Average number of tweets per day before the release date	MGC	Vol
22	Avg($DailyTweets_{1week}$)	Average number of tweets per day 1week before the release date until the release date	MGC	Vol
23-24	N($Reactions$)	Number of replies, retweets before the release date	UGC	Vol
25-26	Pct($Reactions$)	Percentage of replies, retweets before the release date	UGC	Vol
27-28	N($Reactions_{1week}$)	Number of replies, retweets 1 week before the release date until the release date	UGC	Vol
29-30	Pct($Reactions_{1week}$)	Percentage of replies, retweets 1 week before release until the release date	UGC	Vol
31-36	N($SentReactions$)	Number of positive, neutral, negative replies, retweets before the release date	UGC	Comb
37-38	R($Reactions$)	Ratio positive versus negative replies, retweets before the release date	UGC	Val
39-44	Pct($SentReactions$)	Percentage of positive, neutral, negative replies, retweets before the release date	UGC	Val
45-49	N($SentReactions_{1week}$)	Number of positive, neutral, negative replies, retweets 1 week before the release date until the release date	UGC	Comb
50-51	R($Reactions$)	Ratio positive versus negative replies, retweets 1 week before the release date until the release date	UGC	Val
52-57	Pct($SentReactions_{1week}$)	Percentage of positive, neutral, negative replies, retweets 1 week before the release date until the release date	UGC	Val
58-59	Avg($SentReactions$)	Average sentiment of replies, retweets before the release date	UGC	Val
60-61	Avg($SentReactions_{1week}$)	Average sentiment of replies, retweets 1week before the release date until the release date	UGC	Val
61-62	Avg($DailyReactions$)	Average number of daily replies, retweets before the release date	UGC	Vol
63-64	Avg($DailyReactions_{1week}$)	Average number of daily replies, retweets 1 week before the release date until the release date	UGC	Vol

In Table 4, MGC consists of all variables related to the posts and tweets posted by the page administrators themselves. This can again be subdivided in three categories WOM types: volume or valence or a combination of both. First, volume variables are frequency-based variables restricted to a

certain time-window (i.e., the release data of the movie). We follow the recommendation of Ding et al. [24] and Kim et al. [47] and do not only include variables before, but also one week prior to release. Asur and Huberman [5] identify the period 1 week prior as the most critical period since promotional efforts reach their top one week before release and the hype fades out two weeks after release. Since there is no consensus in literature whether to include the absolute or the relative frequency, we implemented both of them [5, 43]. Valence measures are sentiment variables calculated in relation to the release date. We included both the average sentiment score as well as a classification into positive, negative or neutral (see Section 4.3.1). For example, the average sentiment of a post or the percentage of positive tweets before release are unrestricted valence measures. We also included the positive/negative ratio (i.e., the total number of positive posts or tweets divided by the total number of negative posts or tweets) [5]. Finally, combination measures calculate the frequency of sentiment classes. For example, the total number of positive tweets before release is a combination measure. We note that the percentage of positive posts is a valence variable since it represents a relative number, whereas the total number of positive posts is a count variable. Finally, UGC refers to replies and retweets of tweets and comments on posts on the official movie page. These variables are fairly similar to their MGC counterparts. Next, we elaborate on the text analysis and sentiment analysis part of our approach.

4.3.1. Text and sentiment analysis

Since social media posts (especially comments and tweets) are often short informal messages, the content is often cluttered with special characters, typos and emoticons. Since emoticons are often seen as a noisy sentiment label [60], we first transformed the emoticons in the text to their underlying meaning based on an adapted list of emoticons on Wikipedia [82]. For example, :-), :), :-], :] were all coded as happy and :(, :-(, :-C, :C as sad, ;-), ;) as wink and < 3 as heart. Next, we perform the following text cleaning steps. First, we cleaned the text by removing special characters, punctuation, numbers, unnecessary white spaces, stopwords, HTML links, user mentions on Facebook and Twitter, hashtags and retweet entities on Twitter. Second, we ran the text through a spelling-checker based on the Levenshtein distance. Misspelled words were replaced with the most probable alternative based on the Levenshtein similarity index. Third, we applied lemmatization to transform various inflected word forms back to their root form (e.g., driving, drove, driven were transformed to drive). Lemmatization is considered as more accurate than stemming, which only removes word inflections [4]. To perform lemmatization, we use an additional lexicon based on the English lemmatization list of Mechura [59].

After the text was cleaned, we performed sentiment analysis using the lexicon-based method. We chose the lexicon-based approach since it is the only method that scales and affords a fully automated data approach without human interaction. The method does not require maintenance for changing social media and therefore is most attractive for practitioners. The lexicon-based method compares each word in the text-item to a predefined lexicon. If a particular word is located in the lexicon, it assigns the matching valence-score to the focal word [76]. The valence-scores range from $[-4, 4]$, with a

value of 0 reflecting a neutral, -4 a very sad and 4 a happy word. If a word was preceded by a negation, we assigned the opposite sentiment score (e.g., happy was coded as 4 and not happy as -4). We note that emoticons are added to the sentiment score by means of their underlying meaning in the text. The final valence score is achieved by averaging across all the words in the text-item and ranges from -4 to 4 (highly negative to highly positive). If no word in the text-item corresponded to a word in the lexicon, we disregarded the text-item from our sentiment analysis. Our lexicon contains common emotional words. Finally, to improve interpretability we also classified all text-items as negative, neutral or positive [69]. Text-items with an average valence-score between $[-0.5, 0.5]$ were assigned as neutral, higher than 0.5 as positive and lower than -0.5 as negative.

4.4. Prediction algorithms

In total we use 7 prediction algorithms: k-nearest neighbors (KN), decision trees (DT), regularized linear regression (LR), neural networks (NN), bagged trees (BT), random forest (RF), and gradient boosting (GB). We chose these algorithms since they handle different levels of complexity [9] and have proven to yield good performance in sales predictions using social media data [17]. The most fundamental algorithm in box office revenue prediction is *multiple linear regression* (LR) [5]. Multiple linear regression is a parametric technique that assumes a linear relationship between predictors and response [63]. Since this method is prone to overfitting, we apply linear regression with lasso (i.e., least absolute shrinkage and selection operator) [41]. We cross-validate the shrinkage parameter (*nlambda*) in terms of Root Mean Square Error (RMSE) by sequencing over all its values. Nonparametric algorithms are also often employed in box office prediction [47]. The simplest non-parametric method in this study is *k-nearest neighbors* (KN). Compared to other algorithms, KN performs all calculations on the test set [47]. In regression, the prediction of a new sample is the average value of the K nearest neighbors. We iterate over all values from $K = \{1, 2, ..., 150\}$ to determine the optimal K in terms of RMSE. Another simple and popular nonparametric method in movie sales forecasting is *decision trees* (DT) [20]. DTs have the advantage of being simple and interpretable [63]. We use the Classification and Regression Tree (CART) approach by [11] using binary recursive partitioning to build a decision tree. To avoid overfitting we prune our regression trees by cross-validating the cost complexity parameter (*cp*) over all values from $cp = \{0.001, 0.002, 0.003, ..., 0.199, 0.200\}$ and select the value of *cp* that minimizes the RMSE. To cope with the instability and sub-optimal performance of DTs, research in sales forecasting has suggested to use tree-based ensembles [17]. *Bagging* (BT) tries to solve the high variance of decision trees by means of 'bootstrap aggregation' [11]. This implies that independent bootstrap samples of the same size as the training data are constructed by sampling with replacement. For our research we build bagged CART trees with 25 bags (*nbagg*). *Random forest* (RF) adds an additional layer of randomness to the bagging algorithm by only considering a random subset of the predictors at each tree split [12]. Moreover, RF has been identified as the top performing technique in movie success prediction [51]. We set the number of predictors to consider at each split (*mtry*) to the square root of

the number of predictors and the number of trees (*ntree*) to 500 [12]. Besides random forest, boosting algorithms have consistently performed well in social media applications [8]. We use Friedman's *gradient boosting machine* (GB) with CART trees as weak learner to implement the boosting algorithm [36]. GB requires several tuning parameters: the tree depth (*interactiondept*), the number of observations in the terminals node (*nminobsinnode*), the shrinkage parameter (*shrinkage*), and the number of iterations (*n.trees*). We optimize the aforementioned parameters in terms of RMSE by performing a grid search across: $interactiondept = \{1, 3, 5, 7\}$, $ntrees = \{100, 500, 1000\}$, $shrinkage = \{0.01, 0.1\}$, and $ntrees = \{100, 500, 1000\}$ [48, p. 203-208]. Finally, *neural networks* (NN) have also been widely-used in box office prediction [20]. We use a feed-forward neural network optimized by BFGS with one hidden layer and a logistic activation function as our implementation [25]. We choose a NN with one hidden layer since research has shown that one layer NNs are universal approximators of any continuous function [7]. We optimize the weight decay parameter (*decay*) and the number of nodes in the hidden unit (*size*) by performing a grid search across the values for $size = \{5, 10, 20\}$ and $decay = \{0.001, 0.01, 0.1\}$ [66, p. 163 - 170].

4.5. Performance evaluation

We use the root mean squared error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the R^2 to evaluate the performance of the different algorithms. The RMSE, MAE, MAPE, and R^2 are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}_i|, \tag{1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}, \tag{2}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i - \hat{Y}_i|}{Y_i}, \tag{3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = 1 - \frac{SS_E}{SS_T}, \tag{4}$$

with N the number of observations (218), Y_i the actual box office revenue, $\hat{Y}_i$ the predicted box office revenue, and $\bar{Y}$ the mean box office revenue. The SS_E represents the sum of squared errors and the SS_T represents the total sum of squares. In predictive modeling, the R^2 is often calculated as the squared Pearson correlation between the predicted and actual values [48, p. 95].

To make sure our results are robust we employ five times two-fold cross-validation (5x2cv) [23]. This method starts by randomly splitting the data in two equal folds. Each fold gets utilized twice: once as a training set and once as a test set. The whole procedure is repeated five times, which results in 10

performance measures each. We report the median 5x2cv performance measures for each model. To test for significant differences between the various data sources and variable types (see Table 3), wins-ties-losses tables are constructed [21]. To test for significant wins-ties-losses we use the non-parametric Wilcoxon rank test [83]. We also adapt the p-values with Bonferroni-Dunn corrections to control for multiple comparisons and family-wise error [21].

4.6. Information-fusion sensitivity analysis

To uncover which variables are driving predictive performance, we conduct information-fusion sensitivity analysis. Information-fusion is a technique which combines multiple prediction models into one fusion model. This fusion model produces more accurate and reliable results than the individual prediction models [63]. An individual prediction model i with a dependent variable y and n independent variables $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ can be represented as:

$$\hat{y}_i = f_i(x_1, x_2, ..., x_n) = f_i(\mathbf{x}), \tag{5}$$

with $\hat{y}_i$ the predicted response and f_i a certain functional form. The information-fusion model with 7 prediction models can then be represented as:

$$\hat{y}_{fusion} = \Psi\left(\hat{y}_1, \hat{y}_2, ..., \hat{y}_7\right) = \Psi\left(f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_7(\mathbf{x})\right), \tag{6}$$

with $\hat{y}_{fusion}$ the predictions of the information-fusion model and Ψ the fusion operator. In our case we employ a linear fusion operator such that Eq. 6 becomes :

$$\hat{y}_{fusion} = \sum_{i=1}^{7} \omega_i f_i(\mathbf{x}) \quad where \sum_{i=1}^{7} \omega_i = 1. \tag{7}$$

The value of weighting factor ω_i is proportional to the relative predictive performance of prediction model $\hat{y}_i$. Hence, a lower prediction error of a model yielding $\hat{y}_i$ will result in a larger weight ω_i in Eq. 7 and hence more influence in the calculation of the information-fusion model $\hat{y}_{fusion}$. In our case, we calculate ω_i for both MAPE and R^2:

$$\omega_i^{MAPE} = 1 - \frac{MAPE_i}{\sum_{i=1}^{7} MAPE_i}, \tag{8}$$

$$\omega_i^{R^2} = \frac{R_i^2}{\sum_{i=1}^{7} R_i^2}. \tag{9}$$

In a next phase, we conduct sensitivity analysis of the input variables using the information-fusion model. In data mining sensitivity analysis is mostly assessed by means of variable importances [8]. The variable importance of a certain variable j is determined by permuting on that variable and redeploying the prediction model using this permuted variable. The difference in MAPE (R^2) before and

after permutation is the variable importance of variable j. To obtain more reliable and robust estimates for the variable importance, we determine our importance using the information-fusion model. By doing so the information of all prediction models is incorporated [63]. If we rephrase Eq. 7 in terms of importance of variable j with 7 prediction models, this becomes:

$$V_{fusion,j} = \sum_{i=1}^{7} \omega_i V_{i,j}, \tag{10}$$

with $V_{i,j}$ the variable importance of model i and variable j. We calculate $V_{i,j}$ in Eq. 10 for both MAPE and R^2 with respectively ω_i^{MAPE} and $\omega_i^{R^2}$ [5]. Eq. 10 then indicates how much the overall MAPE (R^2) would increase (decrease) if variable j would not be included in the model. To obtain one final sensitivity score, we normalize the sensitivity scores of MAPE and R^2 between $[0, 1]$, and, per variable, we take the average of the normalized sensitivity scores. By doing so we fuse the models on multiple metrics, therefore making the analysis more robust. We note that all of the aforementioned measures are 5x2cv cross-validated. This means that we calculate the median 5x2cv mean increase (decrease) in MAPE (R^2) of the variable in Eq. 10.

5. Results

5.1. Predictive performance

Our research questions are: 'Do social media data increase the predictive performance of box office revenue models over and above traditional movie characteristics?'. If yes, 'which platform (Twitter or Facebook) and which type of data (MGC and UGC) has the most added value in box office sales predictions?'. Table 5 summarizes the average performance and standard deviations of the 7 models proposed in Section 4.2 in terms of RMSE, MAE, MAPE and R^2 across all 7 algorithms. To get more insight into how each model compares to the other models, Table 6 summarizes the wins-ties-losses across all 7 algorithms for each performance measure. For example, the comparison of Movie against Fb:base in terms of R^2 informs us that Movie wins in 1 out of the 7 times from Fb:base and loses in 6 out of the 7 times in absolute numbers. We note that both tables are based on the median 5x2cv results for each performance measure (see Appendix A). In the next paragraphs, we elaborate on the major insights from these analyses.

From the methodological perspective, there are several important observations. A first observation is including social media data over and above movie data always leads to an improvement in predictive performance in almost all cases. For example, Facebook outperforms movie data in terms of RMSE by at least 9.60%[6], by 8.40% in terms of MAE, by 7.85% in terms of MAPE, and by 33.02%[7] in the case

[5] We do not calculate Eq. 10 for MAE, since MAPE is a relative version of MAE, and R^2 is a relative version of the inverted RMSE.

[6] This number is calculated by comparing the performance of Fb:base and Movie: $1 - (2.0168/2.2307) = 0.0960$.

[7] This number is calculated as the increase in performance between Fb:base and Movie: $((0.2916 - 0.2192)/0.2192) = 0.3302$.

Table 5: Average (standard deviation) 5x2cv median RMSE, MAE, MAPE and R^2 across all algorithms

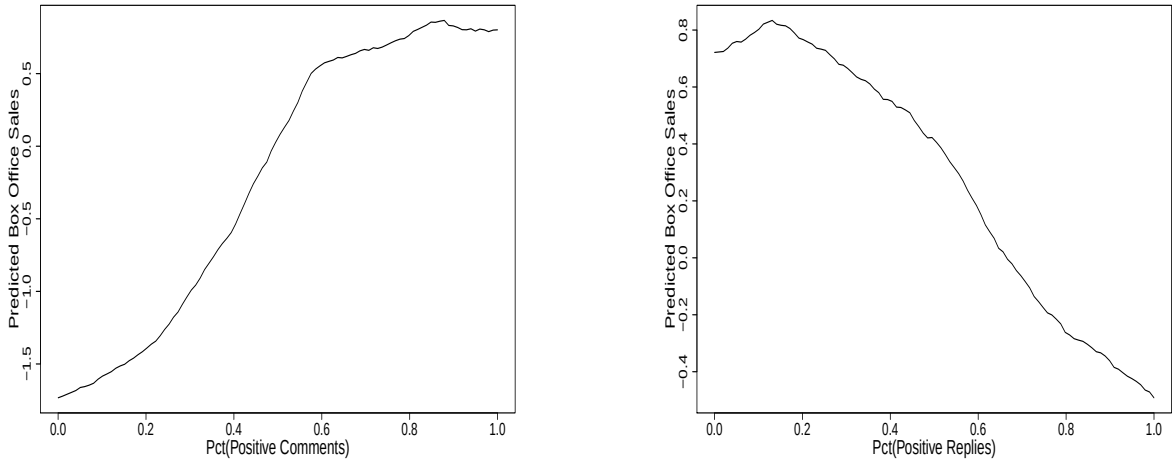
	Movie	Fb:base	Fb:plus	Tw:base	Tw:plus	FbTw:base	FbTw:plus
RMSE	2.2307	2.0168	1.7927	1.9699	2.0481	2.1168	1.8840
	(0.5077)	(0.3789)	(0.2209)	(0.1761)	(0.1859)	(0.3313)	(0.3817)
MAE	1.7624	1.6143	1.3673	1.5233	1.5848	1.6370	1.4186
	(0.4300)	(0.3097)	(0.1838)	(0.1286)	(0.1602)	(0.2865)	(0.2552)
MAPE	0.1172	0.1080	0.0925	0.1094	0.1051	0.1094	0.0952
	(0.0248)	(0.0200)	(0.0113)	(0.0083)	(0.0087)	(0.0162)	(0.0160)
R^2	0.2192	0.2916	0.4408	0.3421	0.2778	0.2624	0.4227
	(0.1113)	(0.0642)	(0.0993)	(0.0877)	(0.0794)	(0.0808)	(0.1349)

of R^2. This increase is always significant for Facebook and mostly for Twitter when UGC is included. Moreover, the combination of Facebook and Twitter always leads to a superior performance compared to traditional movie data, albeit not significant when UGC is not included. A second observation is that the addition of UGC over MGC always leads to a significant increase in predictive performance. A third observation is the superiority of Facebook data over Twitter data across all models. Note that the Tw:plus does perform better than FB:base in absolute terms, however, FB:plus (Fb:base) do significantly outperform Tw:plus (Tw:base). For example, with UGC included Facebook outperforms Twitter in terms of RMSE by 12.47%, by 13.72% in terms of MAE, by 11.99% in terms of MAPE, and by 36.98% in the case of R^2. A final observation is that the Facebook and Twitter model with UGC (FbTw:plus) does not yield a better performance when compared to the Facebook model with UGC (FB:plus). We also see that the FB:plus model is significantly better for 4 out of the 7 algorithms. However, when looking at the results in detail we notice that Fb:plus only performs better for the 4 least performing algorithms. When only looking at the top 3 algorithms (RF, GB, and BT), we see that FbTw:plus significantly outperforms FB:plus. When only taking into account the top 3 algoriths the performance of FbTw:plus becomes 1.5738 for RMSE, 1.1851 for MAE, 0.0799 for MAPE and 0.5535 for R^2. Hence, this might be an indication of overfitting in the case of LR, KN, DT and NN.

From a managerial perspective, we can make the following conclusions. First, social media should always be included if practitioners want to build high performing predictive models. Second, if practitioners want to build models that are both fast and accurate, they should only include Facebook data. The models including Facebook do not have significantly worse performance than models including both Facebook and Twitter. Third, the best overall model includes UGC and MGC data from Facebook and Twitter (FbTW:plus), when looking at the top performing algorithms. However if time is an issue, the Facebook model including UGC (Fb:plus) has equal performance in statistical terms. Nevertheless, UGC should always by included if you want to build the best prediction model.

Table 6: Absolute (significant) wins-ties-losses across all 7 algorithms in terms of RMSE, MAE, MAPE and R^2

Measure	Fb:base	Fb:plus	Tw:base	Tw:plus	FbTw:base	FbTw:plus
RMSE						
Movie	0/0/7 (0/7/0)	1/0/6 (0/2/5)	5/0/2 (0/7/0)	2/0/5 (0/5/2)	1/0/6 (0/7/0)	1/0/6 (0/4/3)
Fb:base	-	1/0/6 (0/2/5)	6/0/1 (2/5/0)	2/0/5 (0/7/0)	5/0/2 (0/7/0)	1/0/6 (0/2/5)
Fb:plus	-	-	7/0/0 (5/2/0)	6/0/1 (5/2/0)	7/0/0 (5/2/0)	4/0/3 (0/7/0)
Tw:base	-	-	-	0/0/7 (0/3/4)	1/0/6 (0/6/1)	0/0/7 (0/2/5)
Tw:plus	-	-	-	-	6/0/1 (1/6/0)	1/0/6 (0/3/4)
FbTw:base	-	-	-	-	-	0/0/7 (0/3/4)
FbTw:plus	-	-	-	-	-	
MAE						
Movie	0/0/7 (0/6/2)	0/0/7 (0/2/5)	5/0/2 (0/7/0)	1/0/6 (0/3/4)	1/0/6 (0/7/0)	0/0/7 (0/1/6)
Fb:base	-	0/0/7 (0/2/5)	6/0/1 (2/5/0)	2/0/5 (0/6/1)	3/0/4 (0/7/0)	1/0/6 (0/1/6)
Fb:plus	-	-	7/0/0 (5/2/0)	6/0/1 (4/3/0)	7/0/0 (5/2/0)	4/0/3 (1/6/0)
Tw:base	-	-	-	0/0/7 (0/2/5)	1/0/6 (0/5/2)	0/0/7 (0/2/5)
Tw:plus	-	-	-	-	5/0/1 (1/6/0)	1/0/6 (0/3/4)
FbTw:base	-	-	-	-	-	0/0/7 (0/4/3)
FbTw:plus	-	-	-	-	-	-
MAPE						
Movie	0/0/7 (0/6/1)	1/0/6 (0/2/5)	6/0/1 (0/7/0)	1/0/6 (0/4/3)	1/0/6 (0/7/0)	1/0/6 (0/1/6)
Fb:base	-	1/0/6 (0/2/5)	6/0/1 (2/5/0)	3/0/4 (0/6/1)	4/0/3 (0/7/0)	2/0/5 (0/4/3)
Fb:plus	-	-	7/0/0 (5/2/0)	5/0/2 (5/2/0)	7/0/0 (4/3/0)	2/0/5 (0/7/0)
Tw:base	-	-	-	0/0/7 (0/3/4)	1/0/6 (0/6/1)	0/0/7 (0/2/5)
Tw:plus	-	-	-	-	6/0/1 (1/6/0)	2/0/5 (0/3/4)
FbTw:base	-	-	-	-	-	0/0/7 (0/4/3)
FbTw:plus	-	-	-	-	-	-
R^2						
Movie	1/0/6 (0/4/3)	0/0/7 (0/2/5)	5/0/2 (2/5/0)	1/0/6 (0/2/5)	1/0/6 (0/7/0)	0/0/7 (0/1/6)
Fb:base	-	0/0/7 (0/2/5)	7/0/0 (4/3/0)	3/0/4 (0/6/1)	5/0/2 (1/6/0)	0/0/7 (0/3/4)
Fb:plus	-	-	7/0/0 (6/1/0)	6/0/1 (4/3/0)	7/0/0 (6/1/0)	3/0/4 (1/6/0)
Tw:base	-	-	-	0/0/7 (0/2/5)	0/0/7 (0/3/4)	0/0/7 (0/2/5)
Tw:plus	-	-	-	-	6/0/1 (2/5/0)	1/0/6 (0/3/4)
FbTw:base	-	-	-	-	-	0/0/7 (0/1/6)
FbTw:plus	-	-	-	-	-	-



(a) Relationship between box office sales and percentage of positive comments on Facebook before release

(b) Relationship between box office sales and percentage of positive replies on Twitter before release

5.1.1. Differences across platform

After having established the superior predictive ability of Facebook data over Twitter data, the following question arises: 'What might be the underlying theoretical mechanisms of this differential predictive performance?'. To model this relationship we need to consider nonlinear relationships. The top performing algorithm in our benchmark is random forest, and is inherently nonlinear. Therefore random forest is a natural choice for this analysis. Random forest, just like any other method is subject to multicollinearity. Multicollinearity is an issue when one is interested in the relationships that govern the model. Of note, this is not an issue when one is solely interested in prediction. Here, we are interested in the relationships and we therefore need to adopt an estimation approach that addresses the multicollinearity issue. We do this in two stages. First, we apply singular value decomposition to the independent variables and extract the singular vectors. Second, these vectors are then used in a random forest model as independent variables. Because the vectors are orthogonal, there is no multicollinearity with respect to their effects.

Because the singular vectors can be expressed as linear combinations of the original independent variables and vice versa, the relationships between the original independent variables and the dependent variable can be computed as a function of the relationships of the singular vectors. This approach is widely known in statistics and marketing [70]. We apply this approach to all variables that were used in the FbTw:plus model. The R^2 of this model amounts to 0.2900 and should be compared to the performance on the original variables, which is 0.4227. Finally, we then visualize the relationship between box office sales and the percentage of positive comments (replies) on Facebook (Twitter) in Figure 2a (2b) using partial dependence plots.

The results confirm our prediction that Twitter displays the too-good-to-be-true effect and Facebook does not. In Figure 2a we see that box office sales increases with the percentage of positive Facebook

comments. In Figure 2b we first note an increase and then a decrease of box office sales with an increase of the percentage of positive replies, similar to what Maslowska et al. [58] find in a different context.

Based on our theoretical work above using Warranty Theory and Signaling Theory this suggests that overall, Twitter has lower source credibility, and coupled with the fact that UGC constitutes the most important predictors in both Facebook and Twitter, this is a plausible explanation of the inferior predictive ability of Twitter. In other words, the results imply that Twitter has less of an impact on box office sales than Facebook because other consumers are less likely to lend credibility to the content generated by other users when they are gathering pre-release information about a new movie. This finding is also reinforced by the network structure of Twitter and the fact that Twitter is mainly used as a platform for entertainment and promotion [13]. For example, if an influencer on Twitter (e.g., a media personality, industry insider or marketer) generates buzz about a certain movie, our theory states that this effect can be predictive, but are negatively related to sales. If a person with an incentive to say positive things says something positive, people will question whatever that person says more than if there is no incentive and people will give less value to that post according to our theory. Also, on Twitter MGC is also often masked as UGC, however if users start to believe that this UGC is manipulated by the marketer, the effect on sales can again be predictive but negative.

5.1.2. Differences across movie type

Another important question is whether or not the predictive power of social media data differs across movie type. To answer this question, we have applied the following procedure. First, we predict box office revenues using only social media data. This means that we included MGC and UGC from both Facebook and Twitter. To model this relationship we used a random forest since this was the best overall algorithm (see Section 5.1). Second, we compare our predictions $\hat{Y}_i$ with the observed box office revenue Y_i and calculate the absolute percentage error (APE) as follows:

$$APE_i = \frac{|Y_i - \hat{Y}_i|}{Y_i} 100. \tag{11}$$

Finally, we grow a CART decision tree [11] with the APE as dependent and the movie types (i.e., traditional movie characteristics) as independent variables. In Table 7 we summarize the predicted APE, the decision rule and the percentage of cases covered by each rule. For example, crime and action movies are the easiest to predict. For other movie genres, the length of the plot, the runtime and whether or not the movie is released in the Summer mainly determine how hard it is to predict box office sales. Movies with a duration between 95 and 100 minutes and no crime movies are the hardest to predict. This information is very valuable for the user of these predictions, and warnings can be issued if a given movie falls into a segment that is hard to make predictions for.

Table 7: Overview of decision rules

APE	Rule	Coverage
2.4	when Crime Movie & Action Movie	10%
3.7	when Crime Movie & Runtime $\geq$ 100 & Length_plot $\in [171, 217]$	14%
4.1	when No Crime Movie & Runtime $<$ 94 & Length_plot $\geq$ 166	12%
5.5	when Crime Movie & No Action Movie	8%
6.1	when No Crime Movie & Runtime $\geq$ 100 & Length_plot $<$ 171 & Summer Season	20%
7.6	when No Crime Movie & Runtime $<$ 94 & Length_plot $<$ 166	6%
10.6	when No Crime Movie & Runtime $\geq$ 100 & Length_plot $<$ 171 & Summer Season	8%
10.9	when No Crime Movie & Runtime $\geq$ 100 & Length_plot $\geq$ 217	8%
11.4	when No Crime Movie & Runtime $\in [94, 100]$	13%

5.2. Sensitivity analysis

Another research question was: 'Which variables are most important?'. More specifically we are interested in which variables from which platform (Movies, Twitter or Facebook) and which data type (MGC or UGC) are important. To do so, we performed information-fusion sensitivity analysis with all Facebook and Twitter variables included (i.e., the FbTw:plus model). The variable importances ($V_{i,j}$) in Eq. 10 are calculated as the 5x2cv median mean increase (decrease) in MAPE (R^2) of permuting variable j in algorithm i, whereas the weights (w_i) are respectively the weighted averages of 5x2cv median MAPEs (R^2s) of the FbTw:plus model (see final column in Table A3 and A4 in Appendix A). Finally, since these sensitivity scores were calculated on different measures (mean increase in MAPE and mean decrease in R^2), we normalized both scores between $[0, 1]$ and took the average to get the final sensitivity scores. Table 8 summarizes the top 20 variables based on information-fusion sensitivity analysis. A detailed description of the variables can be found in Table 4. Next to the rank, the variable and the sensitivity score we also added a column specifying the platform (i.e., Movies, Facebook or Twitter), the data type (i.e., MGC or UGC), and the summary type (i.e., volume (vol), valence (val) or a combination of both (comb)).

First, the results indicate that Facebook is the most important social media platform. A total of 4 variables out of the top 5 variables are related to Facebook and in total 7 out of the 10 variables are from Facebook, whereas 2 from Twitter and 1 from movie characteristics. When looking at the top 25 variables, Facebook is also the most prevalent platform with 44% (= 10/20) of the most important variables related to FB, 32% (= 8/25) related to TW, and 24% related to MOV (= 6/25). This finding confirms the results of Oh et al. [62], namely that Twitter data become less significant when including Facebook data. Second, most of the top predictors are related to UGC (52%), followed by MOV (25%) and MGC (25%). Hence, we confirm the findings of Goh et al. [38] that UGC has a stronger link with movie performance than MGC. This can be explained by consumer engagement behavior (CEB) theory as follows [14]. Interactive engagement, expressed as comments and replies on Twitter,

Table 8: Top 25 variables based on information-fusion sensitivity analysis

Rank	Variable	Sensitivity score	Platform	Data	Type
1	N(*Comments*)	1.0000	FB	UGC	Vol
2	N(*NegativeComments*)	0.8788	FB	UGC	Comb
3	Runtime	0.4027	MOV	-	-
4	Avg(*DailyComments*)	0.3494	FB	UGC	Vol
5	N(*PositiveComments*)	0.3417	FB	UGC	Comb
6	N($PositiveComments_{1week}$)	0.3001	FB	UGC	Comb
7	N(*Replies*)	0.2994	TW	UGC	Vol
8	N(*NeutralComments*)	0.2601	FB	UGC	Comb
9	Pct($NeutralComments1_week$)	0.2521	FB	UGC	Val
10	Pct($Tweets_{1week}$)	0.2512	TW	MGC	Vol
11	Pct($PositiveComments_{1week}$)	0.2502	FB	UGC	Val
12	Pct($Posts_{1week}$)	0.2089	FB	MGC	Vol
13	Avg(*DailyReplies*)	0.2079	TW	UGC	Vol
14	*Comedy*	0.2009	MOV	-	-
15	*Documentary*	0.1994	MOV	-	-
16	Pct($PositiveReplies_{1week}$)	0.1921	TW	UGC	Val
17	N(*PostiveRetweets*)	0.1889	TW	UGC	Comb
18	N(*Retweets*)	0.1838	TW	MGC	Vol
19	N($Comments_{1week}$)	0.1051	FB	UGC	Vol
20	*Rated_PG*13	0.1805	MOV	-	-
21	*Adventure*	0.1768	MOV	-	-
22	Pct(*Tweets*)	0.1657	TW	MGC	Vol
23	*Rated_R*	0.1635	MOV	-	-
24	AVG($SentTweets_{1week}$)	0.1605	TW	MGC	Val
25	N(*Post*)	0.1574	FB	MGC	Vol

Note: FB represents Facebook, TW Twitter, MOV movie characteristics, MGC marketer-generated content, UGC user-generated content, Val valence, Vol volume, Comb combination of volume and valence.

represents engagement in the community and is also clearly of major importance. When we look at social media variables in particular, we find that in terms of incidence in the top 25, volume measures are most important (53% (= 10/19)), followed by a combination of volume and valence (26%) and valence (21%). If we only look at the top 10 variables, we notice that there is only one valence variable present, but the volume and combination measures are equally important. This can be explained by the fact that volume measure and combination measures are often correlated. For example, the number of comments before the release and the number of positive and neutral comments before the release are highly correlated. Hence, when deleting the number of positive and neutral comments, the number of comments before release will become even more important and vice versa. Hence, in general we can say that the awareness effect (volume) is more prevalent than the persuasive effect (valence).

6. Conclusion

In this study we assess the power social media data in predicting box office sales, which platform and data type add more value and why, and which variables are driving the predictive performance. To provide an answer to these questions, we introduce a social media analytical approach consisting of two stages. In the first stage, the predictive performance of several models including Facebook and Twitter data is assessed across 7 algorithms. In the next stage, we apply information-fusion sensitivity analysis to summarize the information of all algorithms and determine the most important variables.

The results indicate that both Facebook and Twitter significantly increase the performance of box office revenue prediction models over and above traditional movie data in terms of RMSE, MAE, MAPE and R^2. We found that Facebook is more indicative of box office sales than Twitter. When comparing both platforms with and without UGC, Facebook models have significantly better performance than Twitter models across all performance measures. Irrespective of the platform MGC substantially improves prediction performance over movie data, and UGC is substantially more predictive than MGC. Moreover, we further investigated the potential underlying mechanisms of differential predictive ability between Twitter and Facebook. Based on Warranting Theory, Signaling Theory and the network structure of both platforms, we hypothesize that Twitter has less source credibility than Facebook, and that user-generated content (UGC) on Twitter should therefore be more likely to be subject to the too-good-to-be-true effect than UGC on Facebook. Based on a nonlinear analysis controlling for multicollinearity, we find support for the too-good-to-be-true effect in Twitter and not in Facebook, implying that Twitter has less source credibility. Finally, our information-fusion sensitivity analysis reveals that volume and valence based combination variables pertaining to Facebook comments (i.e., user-generated content) are the most important variables.

7. Limitations and future research

Our research is limited in that it is possible that there are selection effects in play. Strictly speaking, our results only apply to those movies whose producers chose to operate Facebook and Twitter pages. In an attempt to mitigate this we have looked for movies that do not have social media pages. We would have applied propensity score matching or inverse probability weighting to account for this. However, we were unsuccessful in finding a sufficient number of movies during the time frame (2013-2015) of our analysis that did not have these pages. This finding makes sense given the obvious benefits of having these social media profiles. We did consider matching the movies in our set with movies from before social media existed, but this would introduce other issues, such as differences on other dimensions, which would then result in poor propensity score matching. Given that we were unsuccessful in finding a sufficient number of movies without these profiles from the same period as our data, we do not try to generalize to such movies. Our goal is to determine, in this day and age, which platform is more predictive. Our results only generalize to movies active on social media (both Facebook and Twitter),

Table A1: Median 5x2cv RMSE

	Movie	Fb:base	Fb:plus	Tw:base	Tw:plus	FbTw:base	FbTw:plus
LR	1.8355	1.8206	1.8487	2.0829	1.9872	2.0215	1.8713
RF	1.8610	1.7773	1.5745	1.9929	1.7321	1.8419	1.5633
GBM	2.1781	2.0588	1.6713	2.1910	1.8393	2.0652	1.6034
NN	3.3206	2.8719	2.2404	2.7277	2.1743	2.7962	2.6278
KNN	2.2788	2.1088	1.7973	2.3719	2.1017	2.2331	2.0570
DT	2.1388	1.9682	1.7882	2.2448	2.1403	2.0444	1.9098
BT	2.0018	1.8268	1.6286	1.9945	1.8147	1.8153	1.5546

as that seems to be the prevailing norm. Therefore we are confident that we are capturing a relevant and sufficiently broad representation of the market.

We also acknowledge that it is not observable how movie marketers divide their efforts on maintaining their Facebook versus Twitter contents, and that this can account for differences in predictive ability of these two social media platforms. We are not alone is this. Oh et al. [62], the only other study comparing Facebook and Twitter, acknowledge the same. It would also be reasonable to assume that, on the aggregate, a movie marketer will try to maximize the performance of each social medium. Nevertheless, unfortunately the ability to sort this out stretches beyond the limits of our data.

A direction for future research would be not only to predict final box office revenue, but also to predict opening weekend or opening month box office revenues [24] or even movie success [51]. Since the motivation of this study was to assess the value of social media data in box office sales, we chose the most general measure of box office sales (i.e., final box office gross sales). Another suggestion for future research is to include more social media platforms such as YouTube, Yahoo!Movies or Google trends data. A final interesting avenue for future research would be to set up a randomized controlled experiment to investigate the true causal impact of pre-release social media activity on profitability [67]. For example, this would imply that the creation of a control and a treatment group in which the first group is not exposed to social media activity and the other group is exposed. Moreover, we need to run the experiment on both Facebook and Twitter before and after the release of the movie in order to conduct a difference in differences analysis. Moreover, we would need to conduct this experiment for several movies and measure profitability on an individual level.

Finally, we would like to underscore that, although this study has its shortcomings, we are the first to compare the predictive performance of Facebook and Twitter in box office sales using such an extensive set of algorithms and variables, and to provide theoretical arguments to support our findings. As a result, we believe this study makes a valuable contribution to literature on social media and box office sales from both the methodological and theoretical perspective.

Table A2: Median 5x2cv MAE

	Movie	Fb:base	Fb:plus	Tw:base	Tw:plus	FbTw:base	FbTw:plus
LR	1.4454	1.4201	1.4197	1.5721	1.5234	1.5282	1.4359
RF	1.4541	1.4059	1.2080	1.5579	1.3466	1.3963	1.1660
GBM	1.7008	1.5950	1.2335	1.6887	1.4634	1.5699	1.1951
NN	2.6659	2.2781	1.7250	2.1142	1.6739	2.2080	1.8297
KNN	1.9010	1.6808	1.4416	1.9483	1.6616	1.7995	1.6560
DT	1.6607	1.4841	1.3210	1.7247	1.6004	1.5663	1.4534
BT	1.5092	1.4359	1.2226	1.5520	1.3942	1.3906	1.1943

Table A3: Median 5x2cv MAPE

	Movie	Fb:base	Fb:plus	Tw:base	Tw:plus	FbTw:base	FbTw:plus
LR	0.0973	0.0949	0.0990	0.1064	0.1027	0.1036	0.1004
RF	0.0998	0.0935	0.0829	0.1044	0.0916	0.0951	0.0806
GBM	0.1131	0.1080	0.0856	0.1141	0.0973	0.1062	0.0787
NN	0.1685	0.1508	0.1147	0.1368	0.1127	0.1413	0.1185
KNN	0.1271	0.1106	0.0925	0.1300	0.1107	0.1188	0.1115
DT	0.1119	0.1023	0.0892	0.1169	0.1080	0.1056	0.0958
BT	0.1026	0.0959	0.0834	0.1060	0.0945	0.0949	0.0806

Table A4: Median 5x2cv R^2

	Movie	Fb:base	Fb:plus	Tw:base	Tw:plus	FbTw:base	FbTw:plus
LR	0.3536	0.3464	0.3868	0.2471	0.3439	0.2767	0.3995
RF	0.3444	0.3684	0.5414	0.2886	0.4562	0.3564	0.5767
GBM	0.1684	0.2707	0.4765	0.1760	0.4019	0.2623	0.5502
NN	0.0872	0.1951	0.2506	0.1695	0.2870	0.1998	0.2160
KNN	0.1007	0.2606	0.4618	0.0145	0.2513	0.1250	0.3316
DT	0.1865	0.2504	0.4370	0.1043	0.2325	0.2689	0.3513
BT	0.2934	0.3498	0.5313	0.2678	0.4217	0.3476	0.5339

Appendix A: Median Performance

References

[1] Albuquerque, P., Pavlidis, P., Chatow, U., Chen, K.Y., Jamal, Z., 2012. Evaluating Promotional Activities in an Online Two-Sided Market of User-Generated Content. Marketing Science 31, 406–432.

[2] Apala, K.R., Jose, M., Motnam, S., Chan, C.C., Liszka, K.J., Gregorio, F.d., 2013. Prediction of movies box office performance using social media, in: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1209–1214.

[3] Arias, M., Arratia, A., Xuriguera, R., 2014. Forecasting with Twitter Data. ACM Trans. Intell. Syst. Technol. 5, 8:1–8:24.

[4] Asghar, M.Z., Khan, A., Ahmad, S., Kundi, F.M., 2014. A review of feature extraction in sentiment analysis. Journal of Basic and Applied Scientific Research 4, 181–186.

[5] Asur, S., Huberman, B.A., 2010. Predicting the Future with Social Media, in: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 492–499.

[6] Baek, H., Oh, S., Hee-Dong Yang, Ahn, J., 2017. Electronic word-of-mouth, box office revenue and social media. Electronic Commerce Research and Applications 22, 13–23.

[7] Baesens, B., Gestel, T.V., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J., 2003. Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring. The Journal of the Operational Research Society 54, 627–635.

[8] Bogaert, M., Ballings, M., Bergmans, R., Poel, D.V.d., 2019. Predicting Self-declared Movie Watching Behavior Using Facebook Data and Information-Fusion Sensitivity Analysis. Decision Sciences .

[9] Bogaert, M., Ballings, M., Van den Poel, D., 2016. The added value of Facebook friends data in event attendance prediction. Decision Support Systems 82, 26–34.

[10] BoxOfficeMojo, 2019. Box Office Mojo. URL: `http://www.boxofficemojo.com/`.

[11] Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

[12] Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

[13] Buccafurri, F., Lax, G., Nicolazzo, S., Nocera, A., 2015. Comparing Twitter and Facebook user behavior: Privacy and other aspects. Computers in Human Behavior 52, 87–95.

[14] Calder, B.J., Malthouse, E.C., Schaedel, U., 2009. An Experimental Study of the Relationship between Online Engagement and Advertising Effectiveness. Journal of Interactive Marketing 23, 321–331.

[15] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0 Step-by-step data mining guide.

[16] Chintagunta, P.K., Gopinath, S., Venkataraman, S., 2010. The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets. Marketing Science 29, 944–957.

[17] Cui, R., Gallino, S., Moreno, A., Zhang, D.J., 2018. The Operational Value of Social Media Information. Production and Operations Management 27, 1749–1769.

[18] De Vany, A., Walls, W.D., 1999. Uncertainty in the Movie Industry: Does Star Power Reduce the Terror of the Box Office? Journal of Cultural Economics 23, 285–318.

[19] DeAndrea, D.C., Vendemia, M.A., 2016. How Affiliation Disclosure and Control Over User-Generated Comments Affects Consumer Health Knowledge and Behavior: A Randomized Controlled Experiment of Pharmaceutical Direct-to-Consumer Advertising on Social Media. Journal of Medical Internet Research 18, e189.

[20] Delen, D., Sharda, R., Kumar, P., 2007. Movie Forecast Guru: A Web-based DSS for Hollywood Managers. Decision Support Systems 43, 1151–1170.

[21] Demšar, J., 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. J. Mach. Learn. Res. 7, 1–30.

[22] Dhar, V., Chang, E.A., 2009. Does Chatter Matter? The Impact of User-Generated Content on Music Sales. Journal of Interactive Marketing 23, 300–307.

[23] Dietterich, T.G., 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation 10, 1895–1923.

[24] Ding, C., Cheng, H.K., Duan, Y., Jin, Y., 2017. The power of the "like" button: The impact of social media on box office. Decision Support Systems 94, 77–84.

[25] Dreiseitl, S., Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics 35, 352–359.

[26] Du, J., Xu, H., Huang, X., 2014. Box office prediction based on microblog. Expert Systems with Applications 41, 1680–1689.

[27] Duan, W., Gu, B., Whinston, A.B., 2008. Do online reviews matter? — An empirical investigation of panel data. Decision Support Systems 45, 1007–1016.

[28] El Assady, M., Hafner, D., Hund, M., Jäger, A., Jentner, W., Rohrdantz, C., Fischer, F., Simon, S., Schreck, T., Keim, D.A., 2013. Visual analytics for the prediction of movie rating and box office performance. IEEE VAST Challenge USB Proceedings .

[29] Eliashberg, J., Hui, S.K., Zhang, Z.J., 2014. Assessing box office performance using movie scripts: A kernel-based approach. IEEE Transactions on Knowledge and Data Engineering 26, 2639–2648.

[30] Eliashberg, J., Jonker, J.J., Sawhney, M., 2000. MOVIEMOD: An Implementable Decision-Support System for Prerelease Market Evaluation of Motion Pictures. Marketing Science , 226–243.

[31] Eliashberg, J., Weinberg, C.B., Hui, S.K., 2008. Decision Models for the Movie Industry. Springer Science & Business Media, Boston MA.

[32] Ellison, N.B., Steinfield, C., Lampe, C., 2007. The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites. Journal of Computer-Mediated Communication 12, 1143–1168.

[33] Facebook, 2016. Audience Targeting Options. URL: `https://www.facebook.com/business/help/633474486707199`.

[34] Facebook, 2019a. Company Info Facebook Newsroom. URL: `http://newsroom.fb.com/company-info/`.

[35] Facebook, 2019b. Graph API - Documentation. URL: `https://developers.facebook.com/docs/graph-api/`.

[36] Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics , 1189–1232.

[37] Gaikar, D.D., Marakarkandy, B., Dasgupta, C., 2015. Using Twitter data to predict the performance of Bollywood movies. Industrial Management & Data Systems 115, 1604–1621.

[38] Goh, K.Y., Heng, C.S., Lin, Z., 2013. Social Media Brand Community and Consumer Behavior: Quantifying the Relative Impact of User- and Marketer-Generated Content. Information Systems Research 24, 88–107.

[39] Gong, S., Zhang, J., Zhao, P., Jiang, X., 2017. Tweeting as a Marketing Tool: A Field Experiment in the TV Industry. Journal of Marketing Research 54, 833–850.

[40] Guàrdia-Sebaoun, Rafrafi, A., Guigue, V., Gallinari, P., 2013. Cross-media Sentiment Classification and Application to Box-office Forecasting, in: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, Le Centre de Hautes Etudes Internationales d'Informatique Documentaire, Paris, France, France. pp. 201–208.

[41] Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements Of Statistical Learning. 2 ed., Springer.

[42] Hennig-Thurau, T., Houston, M.B., Sridhar, S., 2006. Can good marketing carry a bad product? Evidence from the motion picture industry. Marketing Letters 17, 205–219.

[43] Hennig-Thurau, T., Wiertz, C., Feldhaus, F., 2014. Does Twitter matter? The impact of microblogging word of mouth on consumers' adoption of new movies. Journal of the Academy of Marketing Science 43, 375–394.

[44] Houston, M.B., Kupfer, A.K., Hennig-Thurau, T., Spann, M., 2018. Pre-release consumer buzz. Journal of the Academy of Marketing Science 46, 338–360.

[45] Jain, V., 2013. Prediction of movie success using sentiment analysis of tweets. The International Journal of Soft

Computing and Software Engineering 3, 308–313.

[46] Kim, S.J., Maslowska, E., Malthouse, E.C., 2018. Understanding the effects of different review features on purchase probability. International Journal of Advertising 37, 29–53.

[47] Kim, T., Hong, J., Kang, P., 2015. Box office forecasting using machine learning algorithms based on SNS data. International Journal of Forecasting 31, 364–390.

[48] Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York, New York, NY.

[49] Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., Kannan, P., 2015. From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior. Journal of Marketing 80, 7–25.

[50] Lampe, C.A., Ellison, N., Steinfield, C., 2007. A Familiar Face(Book): Profile Elements As Signals in an Online Social Network, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA. pp. 435–444.

[51] Lash, M.T., Zhao, K., 2016. Early predictions of movie success: the who, what, and when of profitability. Journal of Management Information Systems 33, 874–903.

[52] Li, X., Hitt, L.M., 2008. Self-Selection and Information Role of Online Product Reviews. Information Systems Research 19, 456–474. Publisher: INFORMS.

[53] Liu, T., Ding, X., Chen, Y., Chen, H., Guo, M., 2014. Predicting movie Box-office revenues by exploiting large-scale social media content. Multimedia Tools and Applications 75, 1509–1528.

[54] Liu, Y., 2006. Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. Journal of Marketing 70, 74–89.

[55] Lo, J.P., 2010. The effectiveness of WOM by using Facebook as an implementation in movie industry. PhD Thesis. Califronia State Univeristy. Sacramento.

[56] Luo, X., Zhang, J., 2013. How Do Consumer Buzz and Traffic in Social Media Marketing Predict the Value of the Firm? Journal of Management Information Systems 30, 213–238.

[57] Malthouse, E.C., Calder, B.J., Kim, S.J., Vandenbosch, M., 2016. Evidence that user-generated content that produces engagement increases purchase behaviours. Journal of Marketing Management 32, 427–444.

[58] Maslowska, E., Malthouse, E.C., Bernritter, S.F., 2017. Too good to be true: the role of online reviews' features in probability to buy. International Journal of Advertising 36, 142–163.

[59] Mechura, M., 2019. Lemmatization list: English (en) [Data file]. URL: `https://michmech.github.io/`.

[60] Meire, M., Ballings, M., Van den Poel, D., 2016. The added value of auxiliary data in sentiment analysis of Facebook posts. Decision Support Systems 89, 98–112.

[61] Meire, M., Hewett, K., Ballings, M., Kumar, V., Van den Poel, D., 2019. The Role of Marketer-Generated Content in Customer Engagement Marketing. Journal of Marketing 83, 21–42.

[62] Oh, C., Roumani, Y., Nwankpa, J.K., Hu, H.F., 2017. Beyond likes and tweets: Consumer engagement behavior and movie box office in social media. Information & Management 54, 25–37.

[63] Oztekin, A., Delen, D., Turkyilmaz, A., Zaim, S., 2013. A machine learning-based usability evaluation method for eLearning systems. Decision Support Systems 56, 63–73.

[64] Oztekin, A., Kizilaslan, R., Freund, S., Iseri, A., 2016. A data analytic approach to forecasting daily stock returns in an emerging market. European Journal of Operational Research 253, 697–710.

[65] Reddy, A.S.S., Kasat, P., Jain, A., 2012. Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining. International Journal of Computer Applications 56.

[66] Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press.

[67] Rishika, R., Kumar, A., Janakiraman, R., Bezawada, R., 2012. The Effect of Customers' Social Media Participation on Customer Visit Frequency and Profitability: An Empirical Investigation. Information Systems Research 24, 108–127.

[68] Ruhrländer, R.P., Boissier, M., Uflacker, M., 2018. Improving Box Office Result Predictions for Movies Using Consumer-Centric Models, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM. pp. 655–664.

[69] Rui, H., Liu, Y., Whinston, A., 2013. Whose and what chatter matters? The effect of tweets on movie sales. Decision Support Systems 55, 863–870.

[70] Rust, R.T., Lemon, K.N., Zeithaml, V.A., 2004. Return on Marketing: Using Customer Equity to Focus Marketing Strategy. Journal of Marketing 68, 109–127.

[71] Shaban, H., 2019. Twitter reveals its daily active user numbers for the first time - The Washington Post. URL: https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/.

[72] Sharda, R., Delen, D., 2006. Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications 30, 243–254.

[73] Smith, A.N., Fischer, E., Yongjian, C., 2012. How Does Brand-related User-generated Content Differ across YouTube, Facebook, and Twitter? Journal of Interactive Marketing 26, 102–113.

[74] Stringhini, G., Kruegel, C., Vigna, G., 2010. Detecting spammers on social networks, in: Proceedings of the 26th annual computer security applications conference, pp. 1–9.

[75] Sul, H.K., Dennis, A.R., Yuan, L.I., 2017. Trading on Twitter: Using Social Media Sentiment to Predict Stock Returns. Decision Sciences 48, 454–488.

[76] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M., 2011. Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 37, 267–307.

[77] Twitter, 2019. Twitter About. URL: https://about.twitter.com/nl/company.

[78] Vendemia, M.A., Bond, R.M., DeAndrea, D.C., 2019. The strategic presentation of user comments affects how political messages are evaluated on social media sites: Evidence for robust effects across party lines. Computers in Human Behavior 91, 279–289.

[79] Viswanathan, V., Malthouse, E.C., Maslowska, E., Hoornaert, S., Van den Poel, D., 2018. Dynamics between social media engagement, firm-generated content, and live and time-shifted TV viewing. Journal of Service Management 29, 378–398.

[80] de Vries, L., Gensler, S., Leeflang, P.S., 2017. Effects of Traditional Advertising and Social Messages on Brand-Building Metrics and Customer Acquisition. Journal of Marketing 81, 1–15.

[81] Walther, J.B., Parks, M.R., 2002. Cues filtered out, cues filtered in. Handbook of interpersonal communication 3, 529–563.

[82] Wikipedia, 2019. List of emoticons. URL: https://en.wikipedia.org/w/index.php?title=List_of_emoticons&oldid=900482232.

[83] Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics bulletin , 80–83.

[84] Wong, F.M.F., Sen, S., Chiang, M., 2012. Why Watching Movie Tweets Won'T Tell the Whole Story?, in: Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks, ACM, New York, NY, USA. pp. 61–66.

[85] Zufryden, F.S., 1996. Linking Advertising to Box Office Performance of New Film Releases: A Marketing Planning Model. Journal of advertising research 36.