Predicting Box-Office Sales using Twitter Activity

Group 1

March 2020

Abstract

HERE WE SHOULD WRITE THE ABSTRACT

Contents

1 Introduction

The power of social influence has been studied widely for a long time and is proven to be powerful (Raven, 1964). People assign value to the opinions of other persons and thereby become influenced or adjust their behaviour. This phenomenon can be seen in many aspects of today's world such as travel behaviour, teenagers' risk perception and impulse buying behaviour (Knoll et al., 2015; Páez & Scott, 2007; Rook & Fisher, 1995). This influence can be done directly by word-of-mouth (WOM) but also through more indirect ways such as posters, commercials, speeches and social media. Of importance in these ways of communication is the way the messages are shaped. Previous studies have shown that communication is key in influencing people's thoughts (Bligh & Hess, 2007; Henry, 2008; Ramage et al., 2010). Because of the rise in social media the past decade we are especially interested in the effect of social media on thought-processes and therefore the question *'How does social media shape the message?'* arises.

Specifically, we investigate the effect of social media on the success of movies. New movies are released frequently and personal opinions about movies, for example through reviews, are shared in large quantities online. In 2017 more than 19,500 reviews were shared on the movie review website Rotten Tomatoes alone (Choueiti et al., 2018). Furthermore, it has been shown that the volume of reviews of movies have a significant effect on their sales numbers (Duan et al., 2008). The results of previous research on the effect of valence, the subjective tone underlying the review, are mixed. On the one hand valence has been shown to have a significant effect on movie sales, even if applied to large-scale social media content (Chakravarty et al., 2010). On the other hand, findings suggest that valence does not drive box office performance (Duan et al., 2008; Liu, 2006).

Yet, these papers mainly researched longer text reviews and did not look at the more informal type of social media reviews communicated through for example Facebook comments or tweets. These types of communication differ from the traditional type of reviews in the sense that features like emojis and gifs or memes can be included and that there are rigid constraints on the number of characters that can be used. Therefore, the effect of these tweets might differ from the effect of the traditional type of reviews. In fact, Rui, Lui and Whinston (2013) found that chatter on Twitter does matter and is dependent on the number of followers and the valence of the tweets.

Yet, since 2013 the landscape of movie-selling has been changing. Namely, consumers have started to discover the benefits of online streaming services such as Netflix, HBO and Videoland, with Netflix being the most popular one (AudienceProject, 2019). Although Netflix was founded in 1998, its revenues exploded only after 2012 when it became increasingly active in countries outside the United States. Since then, the company quadrupled its revenues which resulted in a revenue of more than 20,000 million US$ (Netflix, 2020). This success of Netflix has had various effects in the movie industry. On the one hand, adding a blockbuster movie onto Netflix decreases the rate at which it is pirated (Welter, 2012). Yet on the other hand, movie tickets sales have been decreasing (Parlow & Wagner, 2018). In addition, the social media landscape has been changing a lot since that time as well. Since 2013, there has been an extreme rise in the popularity of Instagram, from 90 million to 1 billion active users (Constine, 2018). Additionally, there has been a drop of 100 million active Twitter users (Leetaru, 2019). Therefore we are interested to see if the current effect of volume and valence regarding tweets about movies is still the same as in the study of Rui, Lui and Whinston (2013).

We add to the existing research by incorporating an influence measurement in our model which makes use of the number of retweets, favourites and replies and thereby refines the indication of the user's influence. We also include a cumulative explanatory variable which captures the prior buzz of the movie, since we expect this prior attention to have a significant effect on the sales. This part of research remains unexplored until now. If our findings provide new perspectives on the effect of tweets on box-office sales, this can have important implications for the management of the movie industry, thereby endorsing the social relevance of our research. We formulate the following research question:

Q: *How can we predict box-office daily sales using Twitter activity?*

To investigate the several aspects of this research question, we formulate the following sub-questions:

Qa: *How can we describe the volume effect of Tweets on box-office performance?*

Qb: *How can we describe the valence effect of Tweets on box-office performance?*

Qc: *What is the effect of a tweet's influence, measured by retweets, favourites and replies?*

Qd: *How does prior buzz effect box-office performance?*

We organize our report as follows. First, we describe our dataset and the collection of our data. Then, in section 3, we formulate our models for analyzing our research question. We present the corresponding results in section 4. Lastly, we draw our conclusions and give recommendations for further research.

2 Data

To conduct our research we gather data on various variables. In this section we elaborate on how we do this and give an overview of the descriptive statistics to gain more insight.

2.1 Dependent Variable

Our dependent variable is the daily movie sales in US and Canada. We choose US and Canada since only for that region daily sales were available. These sales are given in US$. In order to calculate the effect of tweets on total movie sales, we consider all the movies that were released during 2019. We collect these movies and their daily sales through boxofficemojo.com. In total, 908 movies were in theatres in US and Canada in 2019. We reduce this sample for several reasons. First of all, we only select the movies that were shown in at least 975 theatres. We do this since we want to investigate the influence of tweets specifically on movies that are widely accessible. In total, in the US and Canada there are around 6,500 theatres (NATO, 2020; Statistics Canada, 2014). We assume that when the movie is shown in at least 15% of these theatres, it becomes easily accessible. Therefore, we take 975 theatres as a threshold. Of our sample, 177 movies are shown in at least 975 theatres.

Secondly, of these 177 movies we exclude the movies which do not have daily sales available or do not have data on their budget. This is the case for 7 movies. In addition, we remove 43 movies that premiered in 2018 and were still running in 2019. Lastly, we exclude 10 movies which did not start on a Friday. Distributors tend to release their movies on a Friday to take advantage of people's free time on the weekend. Yet, around holidays such as thanksgiving, valentines day, independence day and Christmas movies sometimes release on other days of the week. For the sake of our model we remove these movies. We end up with a sample of N = 117 movies. The descriptive statistics of the gross domestic sales for these movies can be found in table 1. Gross domestic sales is defined as the sum of the daily sales of a movie. Furthermore, we plot the average daily sales per movie in figure 1. The mean of $480,000 is indicated by the

red line. The peaks can be explained as the movie being more than normally popular and thus having extra sales.

Add a new figure

2.2 Explanatory Variables

We use several variables as explanatory variables in our research. We collect these variables through boxofficemojo.com, IMDb and Twitter. The variables which we collect through boxofficemojo.com are the number of days the movie was shown in cinemas; the number of theatres showing the movie and whether the movie is a sequel. We consider a movie to be a sequel when it is a next part of a movie series or when it is based on an earlier released movie, book or series. For example, we define 'The Addams Family', 'Dumbo' and 'Frozen2' to be sequels. Through IMDb we collect the budgets of the movies and in case the movie has a prequel we collect whether this prequel was rated $\geq$ 7.5 out of 10. We assume that when the rating is above a 7.5 movie goers were enthusiastic about the movie and thus are more likely to also visit the sequel.

We obtain twitter data from twitter.com. We scrape this data by using the GetOldTweets3 package for Python version 3.6. Per tweet, the GetOldTweets3 package can provide data on text, date and time, number of retweets, number of favorites, number of replies, hashtags and location. For every movie, we collect the number of tweets, the average number of retweets, the average number of replies, the average number of favourites and the average sentiment across tweets for every day across the period the movie was in theatres. To collect this data we specify hashtags for every movie in our sample. These hashtags are used to select tweets from Twitter to analyse. For most movies we used the movie name as a hashtag to collect the tweets about the movie. Yet, for some movies such as 'Serenity', 'Good Boys' and 'Playing with Fire' this method does not work and a lot of irrelevant tweets pop up. For such movies we added the word 'movie' to the hashtag to make sure the tweets are related to the movie, e.g. #SerenityMovie. We are aware that doing so might result in a lower number of tweets for such movies and thus our estimations of the effect of tweets for these movies might be biased. In addition, Twitter prevents the GetOldTweets3 package from extracting many Twitter data in one go. This causes us to only be able to extract around 400 tweets and their accompanying values in one go. Furthermore, we can only extract tweets in intervals of days, not hours or minutes. Therefore, when more than 400 tweets with the specific hashtag were posted on a day we set the number of tweets on that day to 400+ for that day.

The value of 400 makes a large enough sample for the average sentiment, retweets, favourites

and replies to be unbiased across the movies. For every movie we collect the average sentiment, retweets, favourites and replies on the openingsday in US and Canada for both maximum 400 tweets and maximum 1000 tweets. The resulting plots can be found in figure 2.

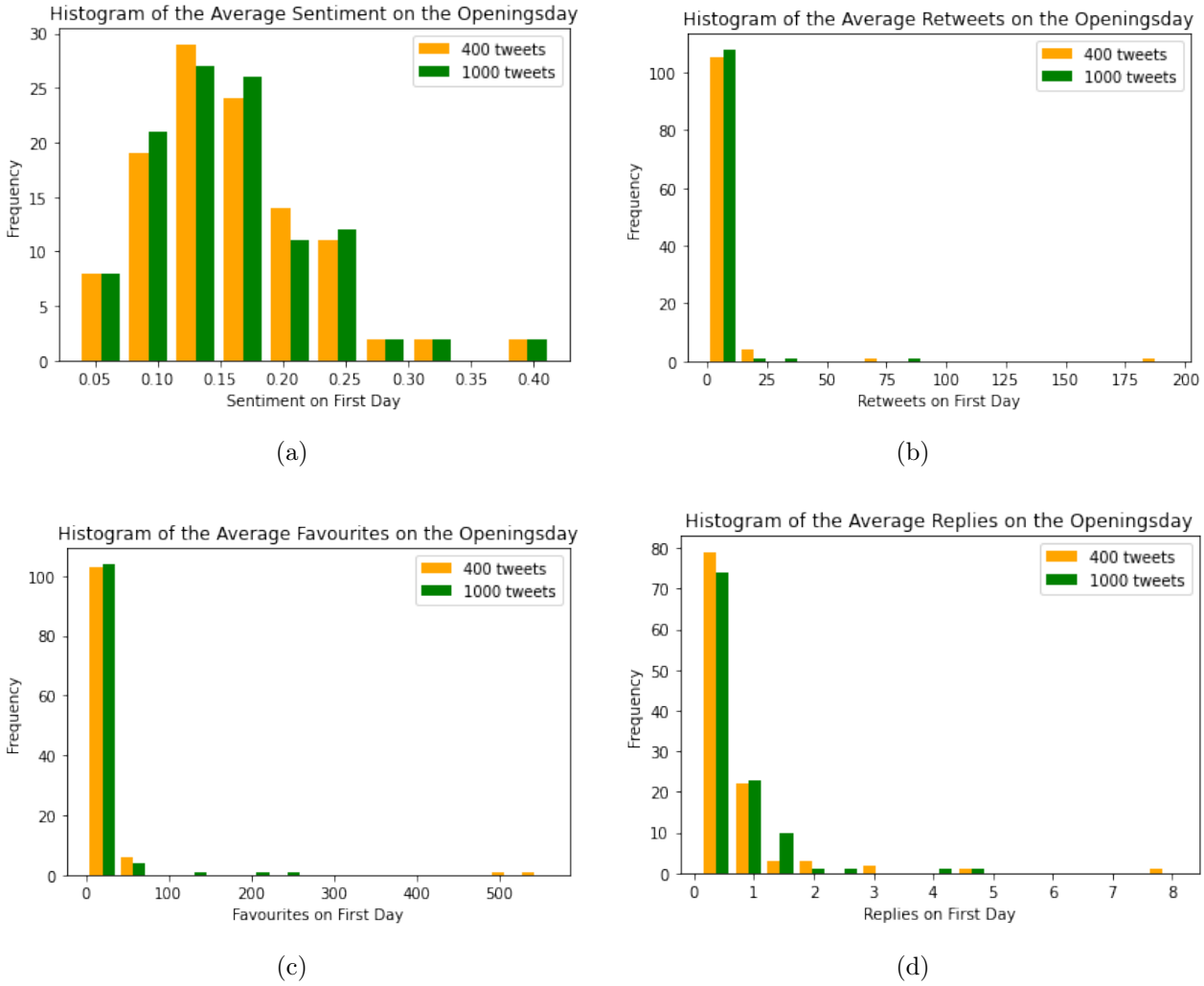


Figure 1: Histogram of Twitter data for number of tweets = 400 and 1000.

From the figure we can see that the histograms for a maximum of 400 tweets are very similar to the histograms for a maximum of 1000 tweets. To formally verify that the distributions with 400 tweets are not significantly different form the distributions with 1000 tweets, we perform the Kolmogorov Smirnov test for each subfigure. The Kolmogorov Smirnov test tests whether two underlying one-dimensional probability distributions differ from each other. The null hypothesis is that the samples are drawn from the same distribution and the alternative hypothesis is that the samples are not drawn from the same distribution. The test statistic is given as $D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$, where$F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of the first and the second sample respectively. The results of the Kolmogorov Smirnov test for average sentiment, retweets, favourites and replies can be found in table 1. From table 1 we can

conclude that for each variable the distribution of the sample with a maximum of 400 tweets is not significantly different from the distribution of the sample where a maximum of 1000 tweets was set in place. Therefore, taking 400 as the maximum number of tweets extracted for one day does not bias our dataset.

Table 1: Kolmogorov Smirnov Test

Variable	Test-Statistic	P-Value
Sentiment	0.045	0.99989
Retweets	0.090	0.76101
Favourites	0.099	0.64926
Replies	0.126	0.34160

The average sentiment across tweets is calculated usin the pre-trained natural language processing model TextBlob. This is a Python library for processing textual data. ..... In this way we are able to get a more insightful sentiment measure than when we would split the sentiment only in the groups 'positive', 'negative' and 'neutral' as done by Rui, Lui and Whinston (2013).

Sentime
Ex-
plain
this
fur-
ther

Lastly we collect the variable PriorBuzz for every movie in our sample. This variable is defined as

$$priorBuzz_i = \frac{\#PriorTweets_i}{\#PriorDays_i}$$

where $\#PriorTweets_i$ is the number of tweets with the specific movie hashtag from the release of the official trailer to the release of the movie and $\#PriorDays_i$ is the number of days between the release of the official trailer and the release of the movie. The dates of the releases of the official trailers can be found on YouTube. Just as for the daily tweet data scraping, there is a limit to the number of tweets that can be collected in one go. In this case this is 6,500 tweets. If the priorbuzz reaches this maximum, we split the priorbuzz period into smaller pieces to be able to get the total number of prior tweets. For some movies like 'The Avengers: Endgame' and 'The Lion King' there were single days for which the maximum was reached. In that case we set the number of prior tweets to 6,500+ for that day, which causes the total number of prior tweet for the 'The Avengers: Endgame' movie to be 124,923+.

Also
talk
about
the
tweets
mavi

The descriptive statistics of these explanatory variables can be found in table 2.

Table 2: Descriptive Statistics

Variable	Mean	St. Dev.	Minimum	Maximum
Movie-Specific Variables				
Gross domestic sales (US$)	73,000,000	117,800,000	1,100,000	858,400,000
Budget (US$)	52,900,000	62,600,000	1,500,000	356,000,000
#Days in Cinema	77	33.34	21	210
#Theatres	3,066	889	1,075	4,802
#PriorTweets (Total)	10,487	18,960	54	124,923
#PriorTweets (English)	7,109	12,097	48	70,314
#PriorDays	109.56	55.44	13	358
Sequel	0.225	0.42	0	1
PrequelRating≥7.5	0.32	0.476	0	1
Movie-Specific, Time-Varying Variables				
Daily Sentiment	0.207	4.509	-0.625	375
#Retweets	483.368	2,200.714	0	81,859
#Tweets	127.35	139.83	0	400
#Favourites	12.381	32.646	0	1,104.992
#Replies	0.517	0.812	0	21.4

Figure 2: Average Daily Sales for all the Movies

3 Methodology

3.1 Predicting Daily Sales in US and Canada

To model the effect of twitter activity on box-office sales in US and Canada we work with a two-stage model. In the first stage we clean the sales data of static effects. Namely, sales of movies can partly be determined by static variables such as budget and prior buzz (Sadikov, Parameswaran & Venetis, 2009; Karniouchina, 2011). Product familiarity also increases the probability of consuming the product, which in the context of movies can be based on consumers' experiences with a prequel (Chakravarty, Liu & Mazumdar, 2010). By cleaning the sales data of these static effects, the estimated effect of twitter activity can be better interpreted since it does not include these static effects anymore.

After obtaining an initial estimate of the daily sales of different movies, which is dependent on the movie-specific static variables, we create a time series to estimate the effect of twitter activity on daily sales. In this time series we include the various twitter variables as exogenous regressors just as done by Rui, Lui and Whinston (2013). In contrast to the study of Rui, Lui and Whinston (2013) we replace the number of followers by the number of retweets as an indication of the tweet's influence. We refine this influence measurement even further by including the number of favourites and the number of replies.

3.1.1 The First Stage: A Movie-Specific model

The first stage of our two-stage model is a regression model for the average sales per day of the movies, μ_i. We estimate this by using movie-specific variables based on data previously explained in the data section.

$$\mu_i = \frac{\text{Gross domestic sales}_i}{\#\text{Days in Cinema}_i} \text{ for } i = 1, ..., N$$

To be precise, these movie-specific variables are the budget of the movie, $budget_i$, the prior buzz of the movie, $priorBuzz_i$ and if a movie is a sequel, D_{1i}, whether or that previous movie was successful based on IMDb rating, D_{2i}.

The dummies are defined as:

$$D_{1i} = \begin{cases} 1 & \text{movie i is a sequel} \\ 0 & \text{otherwise} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{movie i's prequel's IMDb rating} \geq 7.5 \\ 0 & \text{otherwise} \end{cases}$$

Therefore, our first stage equation is as follows:

$$\mu_i = \alpha_0 + \alpha_1 * budget_i + \alpha_3 * priorBuzz_i + \alpha_4 * D_{1i} + \alpha_5 * D_{1i}D_{2i} + \eta_i \text{ for } i = 1, ..., N$$

This is a system of equations with an equation for each movie. We restrict the parameters α_1, α_2, α_3, α_4 and α_5 to be the same for each equation. To adhere to this cross-equation restriction we could use system ordinary least squares (SOLS) to estimate the system. However, using SOLS assumes that there is no endogeneity between the error terms and the regressors within equations. Due to omitted variable bias we believe there might be some endogeneity present. Only using budget, prior buzz and information on prequels is not an exhaustive list of variables that can be used to forecast sales. Keeping this in mind, we use the generalized method of moments (GMM) to estimate the first stage while preserving consistency. We need instrument variables to address the endogeneity concern and to construct the GMM. We define Y = [....], X = [.....] as the endogenous variables and Z = [.....] as the instrument variables. Then the GMM estimator $\beta = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$ can be estimated by solving

$$\min_{\beta}(\sum_{i=1}^{N} Z_i^{\intercal}(Y_i - X_i\beta))^{\intercal} W (\sum_{i=1}^{N} Z_i^{\intercal}(Y_i - X_i\beta))$$

where $W = (N^{-1}\sum_{i=1}^{N} Z_i^{\intercal}\hat{\eta}_i\hat{\eta}_i^{\intercal}Z_i)^{-1}$ is the weighting matrix to produce the GMM estimator with the smallest asymptotic variance and $\hat{\eta}_i$ is a consistent estimator of the residual calculated through System 2 Stage Least Squares (Wang, 2020).

Hausma Test

define instru-ments

add source

3.1.2 The Second Stage: A Time Series model of Weekly Sales per Movie

After obtaining an initial estimate of the daily sales of different movies, which is dependent on movie-specific static variables, we create a time series to estimate the effect of twitter activity on daily sales. We want to analyze over time how daily twitter activity affects movie sales the next day and take into account that sales on one day also affect sales of the same movie later in time. We use the Augmented-Dickey fuller test to confirm that our time series is stable. This leads us to a VAR model with p lags in movie sales. We specify the number of lags by using information criteria. To deal with the fact that movies are released at different points in time, we normalize the release date of every movie to the same date which from now on we will call the 'start date'. This has no effect on the outcome of the model since the absolute start date

Augmen Dickey Fuller Test

Calculat P

is unrelated to our estimation.

The model is formulated as follows:

$$(y_{i,t} - \hat{\mu}_i) = \rho_1 * (y_{i,t-1} - \hat{\mu}_i) + ... + \rho_p * (y_{i,t-p} - \hat{\mu}_i) + \beta_1 * sentiment_{i,t-1}$$
$$+\beta_2 * log(\#type1tweets_{i,t-1}) + \beta_3 * log(\#type2tweets_{i,t-1})$$
$$+\beta_4 * \frac{\#retweets_{i,t-1}}{\#tweets_{i,t-1}} + \beta_5 * \frac{\#favourites_{i,t-1}}{\#tweets_{i,t-1}} + \beta_6 * \frac{\#replies_{i,t-1}}{\#tweets_{i,t-1}} + \beta_7 * D_{5t} + \epsilon_{i,t}$$

As the dependent variable we use the daily sales which are cleaned of the static effects. We achieve this by subtracting $\hat{\mu}_i$, the predicted daily sales, from the actual daily sales for movie i $y_{i,t}$. We regress this on $sentiment_{i,t-1}$, which is the probability of tweets at day t-1 about movie i being positive, $log(\#tweets_{i,t-1})$ which is the logarithm of the number of tweets on day t-1 and three types of influence measures which are the average number of retweets, favourites and replies on day t-1. We take the logarithm of the #tweets because ..... We determine the effect of the number of tweets for two types of tweets. Namely tweets with more than X retweets, we call this type-1 tweets, and tweets with less than X retweets type-2 tweets. In our sample of tweets, 85% had less than X retweets, Therefore, we choose X as the threshold. By dividing the tweets in these two categories we investigate whether tweets that have more retweets and thus are more influential have a different effect on the daily sales than tweets with little retweets.

WHY?

We use $\frac{\#retweets_{i,t-1}}{\#tweets_{i,t-1}}$, $\frac{\#favourites_{i,t-1}}{\#tweets_{i,t-1}}$ and $\frac{\#replies_{i,t-1}}{\#tweets_{i,t-1}}$ to quantify the influence of the twitter activity on day t-1. The more retweets, favourites or replies a tweet has, the more influence it has. We aggregate this variable by taking the average over all tweets during day t. Lastly, we include a weekend dummy variable to account for the fact that more people go to the movies during the weekend.

$$D_{5t} = \begin{cases} 1 & \text{day t is a weekend-day (Friday, Saturday or Sunday)} \\ 0 & \text{otherwise} \end{cases}$$

If the residuals of this model satisfy the white noise conditions, we will estimate the model using OLS. One thing we have to take into account when estimating this model is that the sample sizes of the time series differ across movies. One way to solve this is to fill up the sample of the movies with less daily sales with zeros, also known as 'zero-padding' (Wang and Blostein, 2004). However, this will bias our results because on the days for which we zeropad the sales data there is twitter activity about the movie. In this way we will underestimate the effect of the twitter activity. For that reason we also set the variables of the twitter activity to 0 for days on which the movie was no longer in the cinema.

add source to bibliography

3.2 Top Coding

Table 3: Proportion of Censoring in Total Prior Tweet and Total of Tweets

	Percentage Censored	Total Observations	Not Censored
Total Prior Tweets	13.96%	1,164,115	1,001,615
Total Tweets Per Day	11.26%	782,224	694,145

4 Results

In this section, we present the results of our estimations. We use the general-to-specific method to remove insignificant parameters from our model estimations.

4.1 First Stage

Table 4: Fitted First-Stage Model

Variable	β	(σ)	P-value
C	-0.0267	(0.03)	0.005
budget	-0.0267	(0.03)	0.005
priorBuzz	-0.0267	(0.03)	0.005
prequel	-0.0267	(0.03)	0.005
prequel x rating	-0.0267	(0.03)	0.005

4.2 Second Stage

Maximum lag order to try is seven, this would make a full week (so always including weekends, weekend dummy significant?). In table below we assumed that the lag order for exogenous and endogenous variables are equal. We find an optimal order of p*. As robustness check, we also estimated p*+1 and P*-1 for all tweets, sales combinations. These lead to ... results (Appendix for table).

Table 5: Diebold-Mariano test: fixed λ vs. varying λ

Case	1-month ahead		6-months ahead		12-months ahead	
Maturity (months)	Parameter	P-value	Parameter	P-value	Parameter	P-value
3	-1.9527*	0.0349	-1.3408*	0.0000	1.4888*	0.0000
6	-1.6192	0.1152	0.6966*	0.0000	-0.8067*	0.0000
12	-2.4507*	0.0178	-1.5819*	0.0000	-1.7889*	0.0000
24	-3.5851*	0.0003	-2.8394*	0.0000	-3.0637*	0.0000
36	-3.5991*	0.0002	-2.8741*	0.0000	-3.0631*	0.0000
60	-2.9091*	0.0009	-2.1890*	0.0000	-2.3259*	0.0000
84	-2.2902*	0.0049	-1.5855*	0.0000	-1.6902*	0.0000
120	-1.8971*	0.0138	-1.2118*	0.0000	-1.2957*	0.0000

Table 6: Overview for Lag Order Selection

lag order (p)	AIC	BIC	R^2	$adj.R^2$
p=1	0.9906*	0.9593*	0.9467*	NA
p=2	0.8707	0.6382	0.6399	0.2049
p=3	0.8707	0.6382	0.6399	0.2049
p=4	0.8707	0.6382	0.6399	0.2049
p=5	0.8707	0.6382	0.6399	0.2049
p=6	0.8707	0.6382	0.6399	0.2049
p=7	0.8707	0.6382	0.6399	0.2049

Table 7: F-test Results for Model Selection

model1	model2	F-test result
ALL variables	NO favorites	NaN
ALL variables	NO replies	NaN
ALL variables	NO viral	NaN

Table 8: Robustness Check Lag Order Selection

lag order (p)	AIC	BIC	R^2	$adj.R^2$
p_{tweets} =p*+1	NaN	NaN	NaN	NaN
p_{tweets} =p*−1	NaN	NaN	NaN	NaN
p_{sales} =p*+1	NaN	NaN	NaN	NaN
p_{sales} =p*−1	NaN	NaN	NaN	NaN

Table 9: Fitted Model

Variable	β	(σ)	P-value
$sales_{t-1}$	-0.0267	(0.03)	0.005
$sales_{t-2}$	-0.0267	(0.03)	0.005
$sentiment_{t-1}$	-0.0267	(0.03)	0.005
$\#retweets_{t-1}$	-0.0267	(0.03)	0.005
$\#favorites_{t-1}$	-0.0267	(0.03)	0.005
$\#replies_{t-1}$	-0.0267	(0.03)	0.005
$weekend$	-0.0267	(0.03)	0.005
$viral$	-0.0267	(0.03)	0.005

Table 10: Fitted Model

Maturity	Mean	Std. Dev	Min	Max	MAE	RMSE	$\rho(1)$	$\rho(12)$	$\rho(30)$
3	-0.0267	0.0355	-0.2859	0.0700	0.0244	0.0354	0.7099	0.1981	0.0110
6	0.0391	0.0364	-0.1034	0.1912	0.0265	0.0364	0.6528	0.1703	-0.0028
12	-0.0045	0.0461	-0.1482	0.2542	0.0349	0.0460	0.8338	0.4662	0.1864
24	0.0199	0.0316	-0.0740	0.1077	0.0248	0.0316	0.7816	0.4348	0.2082
36	-0.0230	0.0263	-0.1555	0.0995	0.0197	0.0263	0.6309	-0.0221	-0.0740
60	-0.0214	0.0345	-0.1463	0.0794	0.0256	0.0345	0.8513	0.3269	0.1516
84	0.0252	0.0289	-0.0609	0.1653	0.0211	0.0289	0.8085	0.3496	0.1196
120	-0.0014	0.0296	-0.1311	0.1244	0.0207	0.0296	0.8340	0.3048	0.0015

4.3 Country-Specific Model

5 Robustness

In this section we present several robustness checks. First of all we see how the daily sales are affected by a exogenous shock in one of the regressors through an impulse response analysis

Secondly, we test the predictive performance of our model. Lastly, we include a sensitivity analysis to see what happens when we relax several assumptions.

5.1 Impulse Response

Once the effect of twitter activity is calculated over all movies over the specified time period, we can calculate the effect of a movie-specific exogenous shock at a certain time period, on the sales of tickets for that movie in the future time periods. Therefore, we calculate the impulse response of the movie ticket sales. We investigate the effect of an exogenous shock in daily tweets at time t and an exogenous shock in the sentiment at time t.

We define the exogenous shock in daily tweets as a 5% increase in daily tweets at time t. If there is an exogenous shock on twitter activity at time t, then the effect of this shock on daily ticket sales for that movie at time t+1 is:

$$\begin{aligned}(\hat{y_{i,t+1}} - \hat{\mu_i}) = \rho_1 * (y_{i,t} - \hat{\mu_i}) + ... + \rho_p * (y_{i,t-p} - \hat{\mu_i}) + \beta_1 * sentiment_{i,t} \\ +\beta_2 * (log(\#tweets_{i,t}) + shock)D_{3t} + \beta_3 * (log(\#tweets_{i,t}) + shock)D_{4t} \\ +\beta_4 * (log(\#tweets_{i,t}) + shock)(1 - D_{3t} - D_{4t}) + \beta_5 * \frac{\#retweets_{i,t}}{\#followers_{i,t}} + \beta_6 * D_{5t} + \epsilon_{i,t+1}\end{aligned}$$

The effect of the exogenous shock on twitter activity at time t, carries on to the daily ticket sales for that movie at time t+2 through ticket sales for the movie at time t+1 in the following way:

$$\begin{aligned}(\hat{y_{i,t+2}} - \hat{\mu_i}) = \rho_1 * (y_{i,t+1} - \hat{\mu_i}) + ... + \rho_p * (y_{i,t-p+1} - \hat{\mu_i}) + \beta_1 * sentiment_{i,t+1} \\ +\beta_2 * (log(\#tweets_{i,t+1}))D_{3,t+1} + \beta_3 * (log(\#tweets_{i,t+1}))D_{4,t+1} \\ +\beta_4 * (log(\#tweets_{i,t+1}))(1 - D_{3,t+1} - D_{4,t+1}) + \beta_5 * \frac{\#retweets_{i,t+1}}{\#followers_{i,t+1}} + \beta_6 * D_{5,t+1} \\ +\epsilon_{i,t+2}\end{aligned}$$

Furthermore, we investigate a the effect 5% decrease in average sentiment at time t.

5.2 Prediction

In our sample of movies, some movies were still running in theatres at the time we extracted the data. This was the case for 6 movies. In this section we will first test the predictive performance of our model. We do this by excluding the last 25% of our sample and carrying out an in-sample forecast. Next, we predict the daily sales for our censored movies.

Here we show the predictive performance of our current model

do this

We collected data on the daily sales at March 12 2020. This led to the fact that we have data on some movies which did not go out of theatres yet at March 12 2020. In econometric terms, we are dealing with right censored data. The duration process of the movie being in the cinema has not ended before the end of our observation period. Only counting the observable days the movies were in theatres would lead to biased estimates. Namely, the mean of the variable 'days in cinema' would be underestimated. In addition, our estimates in our time series model could become biased. If we would only include the first observable data of the movies we miss the days with the which is informative and important to our model. To avoid these biases, we exclude these censored movies from our sample. Since these movies only differ from the movies that we do include in our sample for chronological reasons, we assume that they can be modelled in the same way.

To forecast the sales of the remaining days for the censored models non-parametric estimation methods are needed. First, we estimate the expected number of remaining days in theatres through duration methods. The Kaplan-Meier estimator is the standard non-parametric procedure for estimating the distribution function of censored univariate data (...,). We can use this estimate to create a survival function, S(t), a function which models the probability of the object of interest to survive a certain time period (Kaplan & Meier, 1958). The integrated hazard function models the risk of failure, a movie going out of the theatres, at time t. The Nelson-Aalen fitter is a non-parametric estimator of the integrated hazard function and is given by

$$\tilde{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

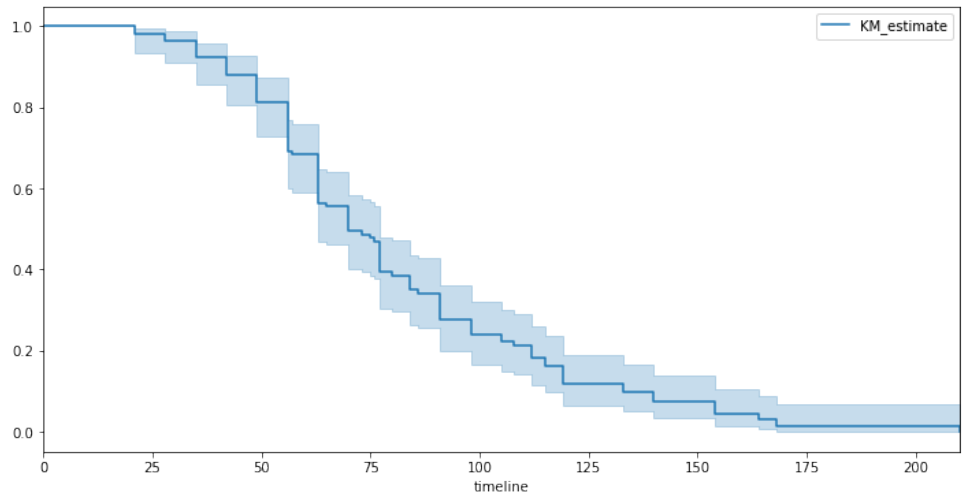
with d_i the number of events at t_i and $n_i n_i the total individuals at risk at t_i t_i$

$$S(t) = 1 - F(t) = Pr[T_i \geq t]$$

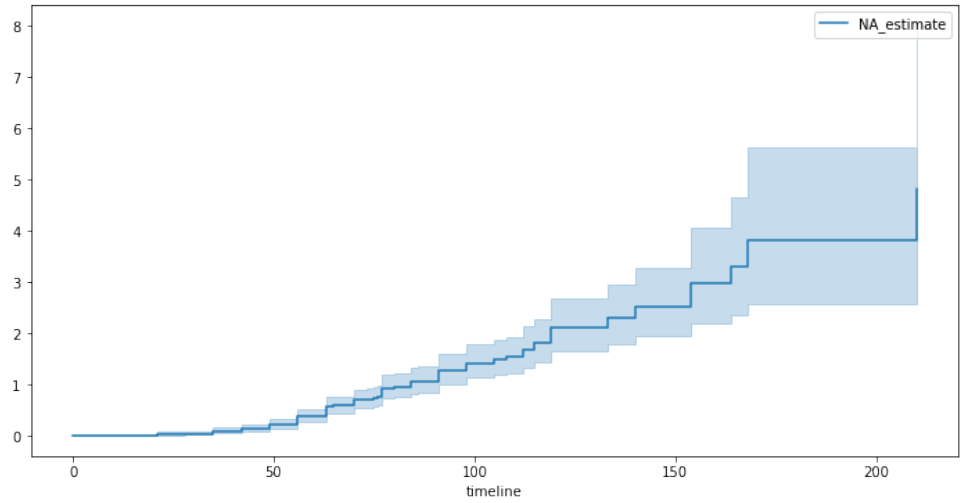
Using the lifelines package in Python we model the baseline survival function and baseline integrated hazard function.

The Kaplan-Meier estimator (Kaplan and Meier 1958) is the standard nonparametric procedure for estimating the distribution function of randomly censored univariate data. → Survival function. S(t) = 1 F(t) = Pr[Ti ≥ t],

Censored regression models are used for data where only the value for the dependent variable is unknown while the values of the independent variables are still available. → we know the index for the days but do not know the daily sales for those days. Also, we can get the



(a)



(b)

Figure 3: Histogram of Twitter data for number of tweets = 400 and 1000.

data for the explanatory variables.

We want to make model in which the predictor is constructed according to information derived from the data → non-parametric estimation methods are needed. This is because we want to avoid biases in our parameter estimates caused by the censoring.

6 Kaplan-Meier estimate + Cox Proportional Hazard model

With the survival function we could estimate the lifetime of the movie. So for the movies that do not have the full daily sales data, we will estimate the day the movie will go out of the cinemas using the survival function. Then for these days we will estimate the daily sales data

using the known sales data of the movie itself and of the other movies.

We experience right censoring. Our time dimension ends before the movies are out of cinemas. S(t) is between zero and one (inclusive), and S(t) is a non-increasing function of t. We will use the Kaplan-Meier Estimate to calculate the survival function for movies being in the cinema. 'The survival probability at any particular time is calculated as the number of subjects surviving divided by the number of people at risk.' Now we use movies instead of people. **In order to see how uncertain we are about the point estimates, we use the confidence intervals**. We will also include explanatory variables to estimate the survival function and estimate the index day when each movie will have ended. We will use Cox Proportional Hazards Model.
To do all this we will use the lifelines package in Python.

Therefore, to avoid biases in our estimates caused by the censoring non-parametric estimation methods are needed. The Kaplan-Meier estimator is the standard nonparametric procedure for estimating the distribution function of censored univariate data (...,). This estimate is used to create a survival function (Kaplan & Meier, 1958).

6.1 Sensitivity analysis

7 Conclusion

8 Bibliography

References

AudienceProject. (2019). *Insights 2019 traditional tv, online video streaming - disney+ special* (tech. rep.). AudienceProject. https://www.audienceproject.com/wp-content/uploads/audienceproject_study_tv_video_streaming_disney_special.pdf

Bligh, M. C., & Hess, G. D. (2007). The power of leading subtly: Alan greenspan, rhetorical leadership, and monetary policy. *The Leadership Quarterly*, *18*(2), 87–104.

Chakravarty, A., Liu, Y., & Mazumdar, T. (2010). The differential effects of online word-of-mouth and critics' reviews on pre-release movie evaluation. *Journal of Interactive Marketing*, *24*(3), 185–197.

Choueiti, M., Smith, S. L., Pieper, K., & Case, A. (2018). Inclusion initiative.

Constine, J. (2018). *Instagram hits 1 billion monthly users, up from 800m in september* (tech. rep. No. 20). Extra Crunch. https://techcrunch.com/2018/06/20/instagram-1-billion-users/

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—an empirical investigation of panel data. *Decision support systems*, *45*(4), 1007–1016.

Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication (1973)*, *45*(4), 363–407.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, *53*(282), 457–481.

Knoll, L. J., Magis-Weinberg, L., Speekenbrink, M., & Blakemore, S.-J. (2015). Social influence on risk perception during adolescence. *Psychological science*, *26*(5), 583–592.

Leetaru, K. (2019). A fading twitter changes its user metrics once again. *Forbes*. https://www.forbes.com/sites/kalevleetaru/2019/04/23/a-fading-twitter-changes-its-user-metrics-once-again/#3a940ffd7a31

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of marketing*, *70*(3), 74–89.

Netflix. (2020). *Results of operations and financial condition.* (tech. rep. No. 21). Netflix. https://d18rn0p25nwr6d.cloudfront.net/CIK-0001065280/b1422088-d225-43b5-b1e9-118daa914ed9.pdf

Páez, A., & Scott, D. M. (2007). Social influence on travel behavior: A simulation example of the decision to telecommute. *Environment and Planning A*, *39*(3), 647–665.

Parlow, A., & Wagner, S. (2018). Netflix and the demand for cinema tickets-an analysis for 19 european countries.

Ramage, D., Dumais, S., & Liebling, D. (2010). Characterizing microblogs with topic models, In *Fourth international aaai conference on weblogs and social media*.

Raven, B. H. (1964). *Social influence and power* (tech. rep.). CALIFORNIA UNIV LOS ANGELES.

Rook, D. W., & Fisher, R. J. (1995). Normative influences on impulsive buying behavior. *Journal of consumer research*, *22*(3), 305–313.

Welter, B. S. (2012). *The netflix effect: Product availability and piracy in the film industry* (Doctoral dissertation). University of Georgia.