FACULTY OF ECONOMICS AND
BUSINESS ADMINISTRATION

Harnessing the Power of Social Media in Predictive Analytics

MATTHIAS BOGAERT

Supervisors: Prof. Dr. Dirk Van den Poel
Prof. Dr. Michel Ballings

A dissertation submitted to Ghent University
in partial fulfillment of the requirements
for the degree of
Doctor of Business Economics

GHENT
UNIVERSITY

Academic year 2017-2018

Typeset in LaTeX.

There is only one thing in the world worse than being talked about, and that is not being talked about.
Oscar Wilde

People leave several traces of their behavior on social media, a key challenge is to find those pieces that provide meaningful insights to firm strategy.

EXAMINATION BOARD

Prof. dr. Patrick Van Kenhove (Dean, Ghent University)

Prof. dr. Dries Benoit (Secretary, Ghent University)

Prof. dr. Bart Larivière (Ghent University)

Prof. dr. Bart Baesens (KU Leuven)

Prof. dr. Koen De Bock (Audencia Business School, France)

Voorwoord

Het schrijven van mijn PhD was niet mogelijk geweest zonder de hulp van verschillende mensen. Daarom zou ik graag enkele mensen extra willen bedanken: vooreerst mijn promotoren Michel Ballings en Dirk Van den Poel. Michel, bedankt voor de vlotte en productieve samenwerking de voorbije jaren. Ik heb steeds met jou op een vlekkeloze manier kunnen samenwerken en ik hoop dat we in de toekomst nog veel projecten succesvol kunnen afronden. Dirk, bedankt om mij de kans te geven om te doctoreren aan de UGent en om steeds in mijn capaciteiten als onderzoeker en lesgever te geloven. Ik apprecieer het enorm dat jij mij de vrijheid hebt gegeven om mijn ding te doen als onderzoeker en mij de lessen 'predictive and prescriptive analyics' hebt toevertrouwd. Je deur stond ook altijd open voor advies en je was er op de cruciale momenten. Ik zou ook graag mijn juryleden (Dries Benoit, Bart Larière, Bart Baesens, Koen De Bock en Patrick Van Kenhove) willen bedanken voor het nalezen van mijn werkstuk en hun constructieve feedback en suggesties.

Ik wil ook graag alle oude en huidige collega's van de vakgroep Marketing aan de UGent bedanken. Of het nu een doordeweekse werkdag, een vakgroepactiveit of een after-work was, de sfeer was altijd opperbest. Ik denk dat vele andere werknemers jaloers zouden zijn op de solidariteit en amicaliteit die er heerst op onze vakgroep. In het bijzonder zou ik toch nog enkele mensen extra in de verf will zetten. Caroline, jij was mijn partner in crime (vooral inzake vettige humor) en je hebt me vaak doorheen moeilijke momenten gesleept. Ik had dan ook enorm moeilijk dat je er het laatste jaar niet meer was, gelukkig heeft Evelynn jouw rol met verve overgenomen. Katrien, we hebben 3 jaar lang naast elkaar vele gesprekken gehad en je kon me steeds motiveren met enkele simpele woorden. Julie, je was de persoon die altijd aan me zag of er iets was en je stond altijd klaar om te luisteren en me gerust te stellen.

Verder wil ik nog mijn vrienden bedanken voor alle leuke avondjes, weekends en reizen. Of het nu een verjaardagsparty, een skireis, een weekendje zee of een boys-night was, jullie gaven me altijd de nodige ontspanning en energie om nadien weer extra hard ertegenaan te gaan. Ik wil ook graag alle spelers ZVC Falcaos en Economini bedanken voor alle mooie sportieve en extra-sportieve momenten.

Tenslotte, zou ik graag mijn ouders willen bedanken voor alle steun tijdens mijn studies en mijn doctoraat. Zonder jullie onvoorwaardelijke steun zou ik nooit staan waar ik nu sta.

Matthias Bogaert

Table of Contents

List of Figures

List of Tables

Nederlandstalige Samenvatting

Sociale media data worden steeds belangrijker in de huidige marketingstrategie van bedrijven. Hiervoor zijn verschillende redenen. Ten eerste is het bewezen dat sociale media activiteit een significante impact heeft op klantengedrag (bv. klanten uitgaven en winstgevendheid). Ten tweede is de hoeveelheid aan sociale media data van ongeziene grootte. Zo heeft Facebook bijvoorbeeld meer dan twee miljard gebruikers - dit komt ongeveer overeen met 25% van de wereldbevolking. Ten laatste, bevat sociale media enorm veel informatie over het gedrag en de eigenschappen van de gebruikers.

Bedrijven die adverteren op sociale media hebben twee mogelijke strategieën: de eerste is een organische strategie. Hierbij proberen bedrijven mond-tot-mond reclame te stimuleren door organisch bereik te kopen en/of door hun sociale media content te optimaliseren. Een andere optie is om te kiezen voor een een-op-een strategie. Hierbij ligt de focus op het identificeren van gebruikers met de grootste kans om jouw product te kopen en deze gespersonaliseerde reclame te sturen. Voor deze een-op-een strategie is het echter nodig om te weten of sociale media data voorspellende waarde heeft. Het doel van dit doctoraat is om de voorspellende kracht van sociale media data na te gaan op verschillende onderzoeksniveaus.

In Hoofdstuk 2 gaan we op gebruikersniveau na of de gegevens van vrienden al dan niet toegevoegde waarde bieden in het voorspellen van de aanwezigheid van de gebruiker op een Facebookevenement. We kunnen besluiten dat dit inderdaad een rol speelt. Daarenboven kan ook het aantal vrienden dat op aanwezig staat, als belangrijke indicator beschouwd worden.

Hoofdstuk 3 focust op het netwerkniveau. Deze studie onderzoekt of gedisaggregeerde interactie-variabelen op Facebook kunnen voorspellen of twee gebruikers een relatie hebben. De resultaten tonen aan dat het mogelijk is om met grote accuraatheid te voorspellen welke gebruikers een koppel zijn. Bovendien tonen we aan dat de gedisaggregeerde variabelen zoals het aantal comments en likes op foto's en video's hierin zeer belangrijk zijn om dit te voorspellen.

Hoofdstuk 4 speelt zich af op het meest geaggregeerde niveau, namelijk productperformantie. Deze studie onderzoekt welk sociale media platform (Facebook of Twitter) de beste voorspeller is van de verkoopcijfers van een film. De resultaten tonen aan dat Facebook een significant betere voorspeller is dan Twitter. Ook tonen we aan dat de content, gegenereerd door gebruikers, geen extra voorspellende waarde heeft tegenover content van het bedrijf of populariteitmaatstaven van de Facebook- en Twitterpagina.

Summary

Social media data are becoming increasingly central to firms' efforts to understand buyers and develop effective marketing strategies. The reasons are manifold. First, social media buzz has proven to have a significant impact on key customer metrics, such as customer spending, cross-buying, and profitability. Second, the volume of social media data is unprecedented. For example, Facebook has more than 2 billion users, corresponding to a staggering 25% of the world population. Finally, social media data contain a lot of information about the preferences and the characteristics of the users.

Companies that want to advertise on social media can adopt two main strategies. The first one is an organic strategy. This implies that companies try to stimulate word-of-mouth by paying for more organic reach and/or by optimizing their social media content. Another option is to choose for a one-to-one strategy. This strategy focuses on identifying the users who are most likely to buy your product and target them directly with personalized ads. In order to implement such an one-to-one strategy, it is important to know whether social media data have predictive value. The goal of this dissertation is thus to harness the predictive capacity of social data on different levels of analysis.

Chapter 2 investigates on the user level whether Facebook friends data have added value in event attendance prediction. The findings show that Facebook friends data significantly improve event attendance models in a majority of the cases. Moreover, we find that the number of friends that attend the event is one of the top indicators of event attendance.

Chapter 3 focuses on the network level. This study investigates whether disaggregated variables can predict romantic partnership on Facebook. The results reveal that it is possible to predict somebody's significant other with high predictive accuracy. We also show that disaggregated variables, such as comments and likes on photos and videos, are among the top predictors of romantic partnership.

Chapter 4 is situated on the most aggregate level, namely product performance. This chapter studies which social media platform (Facebook or Twitter) is the most predictive of movie sales. The results indicate that Facebook is significantly more indicative of movie sales than Twitter. The results also show that user-generated content does not significantly increase the predictive power of models based on marketer-generated content and page popularity indicators of the Facebook and Twitter page.

1 General Introduction

This is a PhD dissertation by publication, this means that the manuscript is a collection of research papers (i.e., published or working papers) that are meant to be standalone. This implies that the chapters can be read independently from each other.

The first chapter introduces the reader with the context, motivations and the overarching goal of this dissertation. It also discusses the holistic analytical framework that is used throughout the different chapters. The third section of this chapter describes the data. The final section is dedicated to the main findings, contributions and practical implications of each study. The goal of this chapter is to provide context for each of the studies in this dissertation. After reading this chapter, the reader should be able to situate the different chapters in the *social media analytics* literature.

1.1 Introduction

Nowadays social media allows companies to gather data in an inexpensive way and on a large scale. For example, Facebook has over two billion users and every second five new profiles are created [84]. This implies that companies can tap into a staggering 25% of the world population. Moreover, social media sites also contain a lot of variables related to a user's behavior and characteristics that cannot be found in traditional databases [136]. This large number of potential observations and the richness in terms of variables results in unparalleled advertising

opportunities for companies. As a matter of fact, a growing number of studies is investigating how firms can use this vast amount of information as a source of insights to develop effective marketing strategies [90].

There are two main advertising strategies that firms can adopt on social media (Figure 1.1): an organic or a one-to-one strategy. The former strategy focusses on increasing the overall advertising awareness and reach on Facebook [23]. To do so firms can focus on their owned and earned social media. Owned social media refers to the social media content that are created and owned by the firms itself. Typically these studies focus on adapting the characteristics of their social media content (e.g., length, type, and timing) to increase reach (e.g., more likes or comments) [61]. Earned social media refers to social media content that is generated by the users in the company's network [222]. These studies typically focus on the impact of user-generated content on firm metrics [60]. However, a general problem with most of these studies is that they often study owned and earned media in isolation and tend to focus on a single social media channel.

A one-to-one strategy implies that firms target the individual users in their network directly to attract their attention [38]. Nowadays targeting on social media is done indirectly via organic strategies. For example, when posting content on Facebook advertisers can decide to target consumers based on their socio-demographic characteristics (e.g., age, gender, and education), location (e.g., country, state or city), interests and behavior (e.g., leisure activities an opinions) [86]. A first problem is that these targeting options are more descriptive in nature instead of predictive. This means that the targeting models have a limited number of features and are rather easy to estimate. A second problem is that the social media platform often limits the organic reach of your network. For example, on Facebook the organic reach has declined to 6.5% of your total network, and even to 2% if your page has more than 500,000 followers [158]. To increase their organic reach advertisers can pay the social media platform to boost their post and propagate it to a larger audience. Hence, firms that want to reach a large audience are almost forced to pay for publicity. A third problem is that the data required to implement this targeting strategy is not readily available. Hence, firms should know the characteristics of the users who might be buy their product (or service). This implies that firms have to conduct in-house marketing research projects or hire market research companies before they can target users on Facebook. In the end, even if they know the characteristics of their customer base, they will still have to pay the social media platform to increase their organic reach. As a result companies might end up paying twice: once to the social media platform and once to the market research company. While this strategy is effective, there is still room for improvement. The underlying reason is that these targeting option and organic strategies are still high-level and the targeting models are rather simple in nature. Hence, we call these targeting options descriptive. For example, these options describe the gen-

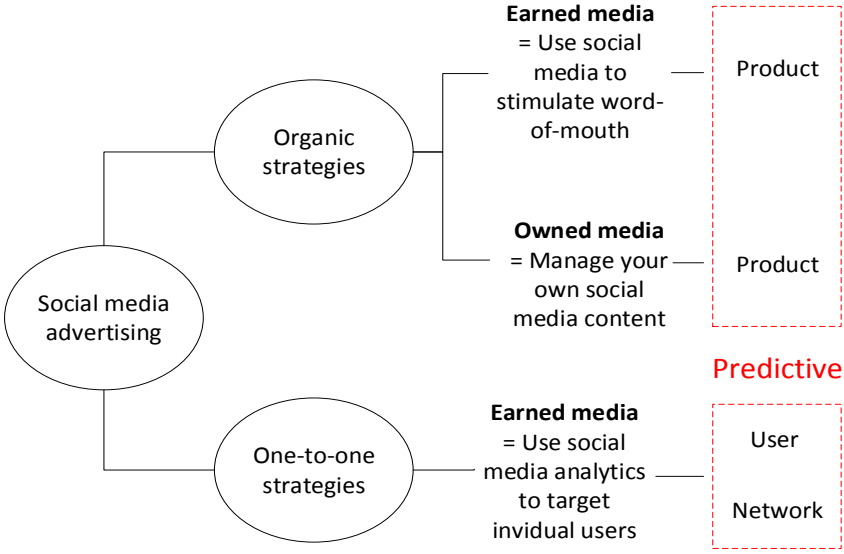


Figure 1.1: Overview of social media advertising options

eral characteristics of the customer base that companies can use target users on. However, companies are still not able to target the user directly. In order to effectively implement an one-to-one advertising strategy, it is necessary to transform these descriptive targeting options to predictive targeting options. For example, predictive targeting implies that a firm would analyze the data from all the users in their network and target those with the highest probability to buy a product. To implement such a strategy it is important to know whether it is feasible to use social media for predictive purposes. Moreover these targeting models should be based on a huge set of candidate features and complex prediction models.

This dissertation contributes to literature by assessing the predictive power of social media for both one-to-one strategies and organic strategies. For one-to-one strategies the most granular level is the user. On this level, we investigate whether or not social media data are able to predict individual user behavior with a user's characteristics. For example, Chapter 2 investigates whether it is possible to predict if a user will attend the focal event using his/her Facebook data. The second level focuses on the network and investigates whether we can predict the relationship between ego and alter using their interactions within the network. For example, Chapter 3 predicts whether or not ego and alter are each other's significant other using disaggregated features based upon the interactions between

ego and alter. The question of whether or not social media are predictive is non-trivial, since it allows companies to implement targeted and proactive marketing approaches [38]. For example, applied to event attendance a targeted marketing approach does not rely on mass advertising to create awareness for the focal event. Instead, targeted strategies try to identify the people with the highest (or smallest) likelihood to attend the event and attempts to appeal to those users with specific actions. A proactive approach in this case involves identifying in advance the invitees who are most likely to attend or not and taking specific action in advance of the event to influence the overall attendance such as sending targeted communications to that person's friends to attempt to influence their behavior. For the organic reach, we focus on whether or not product (or firm) performance indicators can be predicted using data from a product's (or firm's) official social media page. To investigate the predictive capacity of social media, we focus on both users-generated and marketer-generated content from several social media channels. For example, Chapter 4 compares which social media platform is most indicative of box office sales. To evaluate the predictive power of social media, we propose a data analytical system. The purpose of this system is to (1) assess the predictive capacity of social media and evaluate which algorithms perform best, and (2) to determine which variables are the driving force of predictive performance.

The next section introduces the general data analytical framework that is applied throughout each chapter. The final section summarizes the main findings, contributions and managerial implications of three studies.

1.2 Analytical framework

We use an adaptation of the popular CRISP-DM methodology, which stands for 'Cross-Industry Standard Process for Data Mining'. According to a survey on the popular data science community *KDNuggets*, CRISP-DM is the most widely used and well-known methodology in data analytics [183]. CRISP-DM organizes the data mining process in several sequential steps. These steps help practitioners in conducting and structuring the data mining process. Hence, it can be seen as a blueprint for planning and conducting data analytical research [43]. The original CRISP-DM model involves the following six steps:

1. *Business understanding*: This phase involves analyzing the business environment, defining the business objectives and setting the data mining goals to solve the current business problem.

2. *Data understanding*: This step begins with collecting the initial sources and getting familiar with the data. Next, data description and exploration report can be made and possible problems with the data quality can be identified.

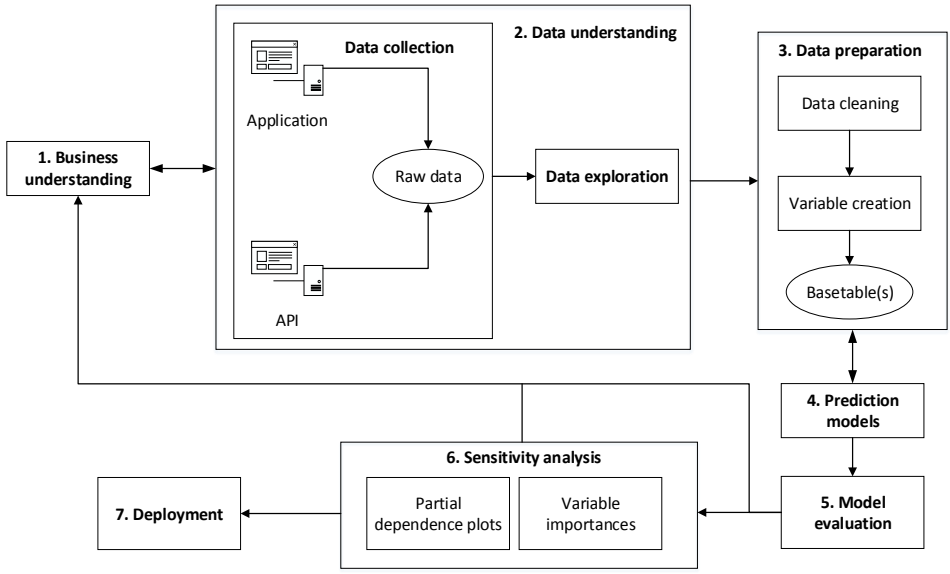


Figure 1.2: Analytical framework

3. *Data preparation*: This step selects and cleans the relevant data sources. Next, various predictors are calculated and the different data sources are merged into the final basetable(s).

4. *Modeling*: This stage selects several prediction algorithms and uses these algorithms to build predictive models. This stage also includes the test design (e.g., cross-validation).

5. *Evaluation*: This steps evaluates the performance of the predictive models developed in the previous step. The results are also checked for consistency with data mining goals and business objectives.

6. *Deployment*: This stage involves deploying the data mining models in real-time to assist managers in the decision making process.

Figure 1.2 summarizes the holistic analytical framework applied in our studies. Several small adaptations are made to this framework. The main modification in our framework is the addition of the *sensitivity analysis* step (step 6 in our

framework). This step assesses which variables are important and uncovers the relationship between predictors and response by means of variable importances and partial plots. As with the *evaluation* step, the results of the sensitivity analysis are tested against the objectives and the goals of the study. One might argue that the sensitivity analysis step can be seen as a part of the *evaluation* phase of the CRISP-DM model. However, we believe that a separate step in the analytical framework is necessary for the following reasons. First, in the original CRISP-DM framework Chapman et al. [43] only implicitly assume that sensitivity analysis is part of the *evaluation* phase. They state that this phase evaluates the accuracy ad the generality of the model. Second, the *sensitivity analysis* step uses the results of both the *modeling* and the *evaluation* step. Hence, it can be seen a synthesis of both steps. Third, the *sensitivity analysis* step requires the modeler to conduct several analyses that are not performed in the *modeling* and the *sensitivity analysis* phase. For example, a fusion model (i.e., hybrid ensemble) of all the individual algorithms is built and/or variable importances and partial plots are constructed. Since this steps is crucial in our approach and it requires a lot additional calculations, we decided to include this as an additional step in the CRISP-DM process to substantiate our contributions. The *data understanding* phase differs from the original CRISP-DM model in the fact that our data sources are gathered via social network sites. This can be done via directly communicating with the API or by developing a customized application that interacts with the API. The *data preparation* step calculates the independent variables. Depending on the application this also requires text mining and sentiment analysis next to the traditional frequency and time-related variables. The *modeling* phase uses several prediction algorithms ranging from statistical parametric models to non-parametric machine learning algorithms and from classification to regression techniques. Next to determining the algorithms, this step also includes the choice of the cross-validation method. The *evaluation* phase employs different performance evaluation metrics depending on the nature of the dependent variable (binary or continuous). Table 1.1 summarizes the methodology of all three studies with respect to the different steps in the analytical framework.

Table 1.1: Overview of the methodology in the three studies

Study	Data understanding	Data preparation	Modeling	Evaluation	Sensitivity analysis
The Added Value Of Facebook Friends Data in Event Attendance Prediction	Customized Facebook app	*Response*: declared event attendance. *Predictors*: Time and frequency variables	LoR, NB, NN, AB, and RF	AUC	Mean decrease in Gini Index and partial plots
Evaluating the Importance of Different Communication Types in Romantic Tie Prediction on Social Media	Customized Facebook app	*Response*: declared significant other. *Predictors*: Time and frequency variables	KN, LoR, NN, RF, AB, KF, and RoF	AUC, accuracy, G-mean, and F-measure	Information-fusion sensitivity analysis and partial plots
Comparing the Ability of Twitter and Facebook Data to Predict Box Office Sales	API	*Response*: gross box office revenues ($). *Predictors*: Time, frequency, text and sentiment variables	LiR, KN, DT, NN, BT, RF, and GB	RMSE, MAE, MAPE, and R^2	Information-fusion sensitivity analysis

Note: AB = adaboost, BT = bagged trees, DT = decision trees, KF = kernel factory, KN = k-nearest neighbors, LoR = logistic regression, LiR = linear regression, NB = naive Bayes, NN = neural networks, GB = gradient boosting, RF = random forest, RoF = rotation forest

1.3 Data

Chapter 2 and 3. To gather our social media data (i.e., Facebook), we created a customized Facebook application for a European soccer team. The Facebook application had a back-end and front-end. The former comprised of several databases to store the collected data. The latter comprised of the features visible to the users. To create awareness and visibility the link to the application was promoted several times on the Facebook page of the European soccer team and added to the main page tabs. To stimulate participation, we offered a signed jersey as an incentive. When the users opened the application they were confronted with an authorization box. This authorization box asked permission to the users to gather their data in exchange for entering the drawing of the prize. Next to asking permission the authorization box also included a rules and regulations section, containing our contact information. The rules and regulations included a list of the collected data and stated that we would only use their data for academic purposes. Afterwards the users had to fill out several questions regarding the soccer team. A question regarding the number of participants of the application would determine the winner of the signed jersey. The data were gathered between May 7, 2014 and June 9, 2014. In total we collected data from 5010 unique users and 1,103,212 friends. Figure 1.3 gives an example of which information was extracted by our application on the Facebook profile of Marc Zuckerberg.

Chapter 4. To extract information from Facebook and Twitter movie pages we used the publicly available API [85, 211]. The Facebook and Twitter API are easily accessible and allow for fast and easy processing of the extracted files [161]. In total we collected data from 231 movies released between January 2012 and December 2015 from their respective Facebook and Twitter pages. We collected the data from the start of their very existence of their Facebook and Twitter page until the time of collection (August 2016). The reason that we only included movies released until December 2015 is to make sure that the movies were out of theaters and thus reached their final gross box office revenues. We collected the box office sales figures via BoxOfficeMojo within the same time window [33]. Figure 1.4 gives an overview of the extracted data from the Twitter (left) and Facebook (right) movie pages. The collected data can be categorized into page-popularity indicator (PPI), marketer-generated content (MGC), and user-generated content (UGC). PPI (green boxes) are indicators of the overall popularity of a Facebook or Twitter page, such as the total number of page likes on Facebook and the number of followers on Twitter [173]. MGC (blue boxes) refers to a movie producer's owned social media, such as the number of Facebook posts and number of tweets created by the page owners [222]. Finally, UGC (red boxes) refers to the earned social media (i.e., the content on the Facebook wall or Twitter pages created by the other users), such as comments on Facebook or replies on Twitter [222].

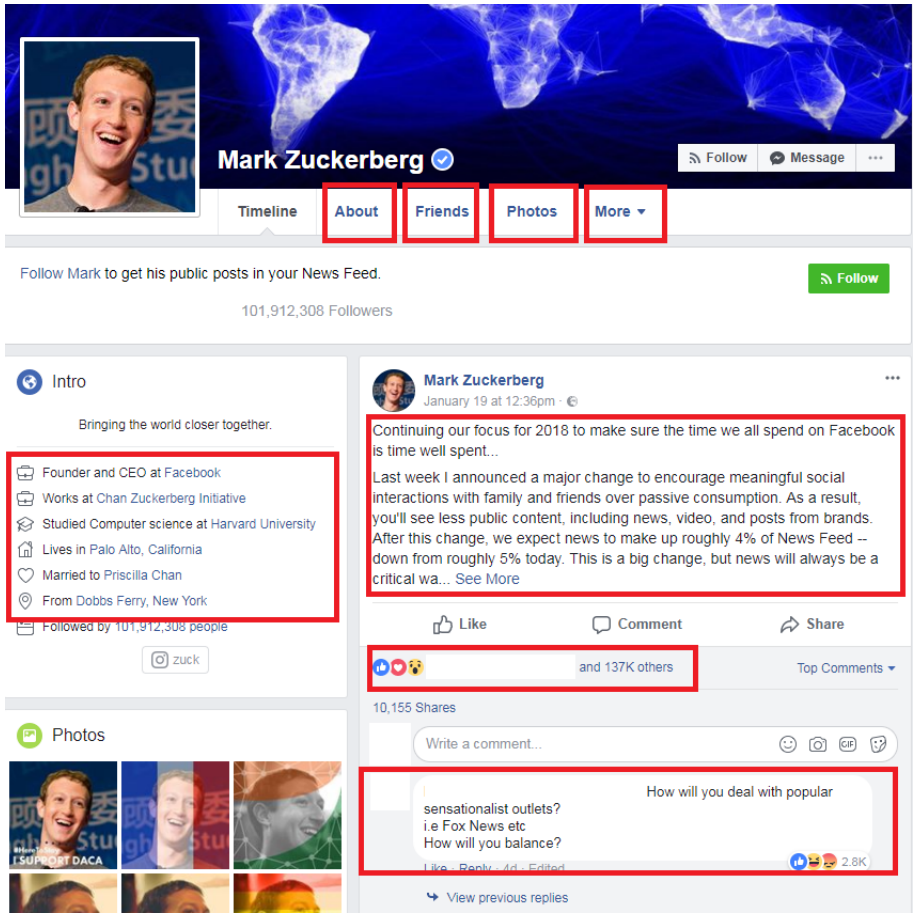


Figure 1.3: Example of the collected data (red boxes) from a user profile in Chapter 2 and 3

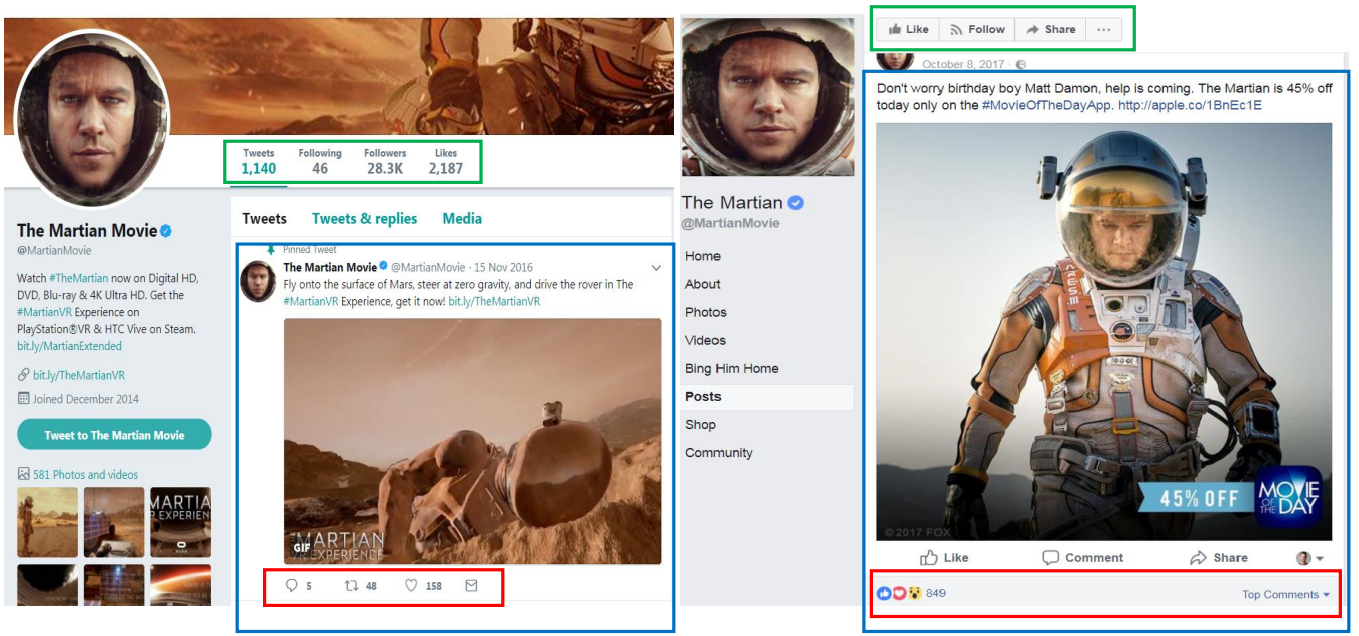


Figure 1.4: Example of the collected data from a Twitter (left) and Facebook (right) page in Chapter 4. The green boxes represent page-popularity indicators, the blue boxes marketer-generated content, and the red boxes user-generated content.

1.4 Extended abstract

Table 1.2 summarizes the contributions, main findings and implications of each study. In the following paragraphs we further elaborate on the findings in each chapter.

Chapter 2. This study aims to (1) evaluate the added value of Facebook friends data over and above user data in event attendance prediction, and (2) identify which variables are most important in event attendance prediction and uncover their relationship with the propensity of attending an event. To evaluate the added value of a user's friend data, we build and compare 2 models across 5 prediction algorithms (i.e., logistic regression, naive Bayes, neural networks, random forest, and adaboost). The first model (i.e., the baseline model) only includes user data, whereas the second model augments the first model with friends data (i.e., the augmented model). To make sure that our results are robust we employed five times two-fold cross-validation. To substantiate whether or not the added value of friends data is significant we use the Wilcoxon signed rank test. The results show that augmenting user data with a user's friend data increases the five times two-fold cross-validated AUC from 0.22%-points to 0.82%-points. The results of our significance test indicate that the increase in predictive performance is significant in three out of the five algorithms, marginally significant for one algorithm, and not significant for one algorithm. In terms of top performing algorithms we find that adaboost is the top performer, followed by random forest, logistic regression, neural networks and naive Bayes for both the baseline and the augmented model. The variable importances show that the absolute and relative number of friends that are attending the focal event are amongst the top predictors. Other important variables are related to the timing of the event, such as the start and the end day of the event. Our findings provide important implications for the Facebook company, event promoters, and companies that want to estimate event attendance. For Facebook Inc., the results suggest that friends data is significant when promoting event. Hence, friends data should be incorporated the News Feed Algorithm when promoting events to users and to increase user experience. Future research can pinpoint whether friends data should be included in other applications. For event schedulers our findings indicate that adding explicit information about the attendees and the timing of the event could increase the probability of attending the event. Finally, practitioners or companies that want to estimate event attendance benefit from our results since we show that adding friends data significantly improves the accuracy of predictive models in a majority of the cases.

Chapter 3. This study assesses which communication types are most important in romantic tie prediction. In contrast to previous studies we do not use aggregated communication features to estimate romantic ties. Instead, we model social ties using disaggregated time and frequency measures. We define disaggregated

features as separate measures for comments, likes, and tags an ego has placed on an alters' statuses, photos, albums, check-ins, and location updates. Next to disaggregated features, we also include socio-demographic and preference features. To ensure robustness of the results we benchmark four single classifiers (i.e., k-nearest neigbors, logistic regresson, naive Bayes, and neural networks) and four ensemble classifiers (i.e., random forest, adaboost, rotation forest, and kernel factory) using five times two-fold cross-validation. Our response variable of interest is whether or not ego is the significant other of alter and vice versa. To cope with the high class imbalance in our binary response variable we employ several sampling techniques (i.e., random oversampling, random undersampling, and synthetic minority oversampling). Finally, we use information-fusion sensitivity analysis to gain insight in the top predictors of romantic ties and the underlying relationship with the response variable. The results indicate that we can predict romantic ties with very high predictive performance with a cross-validated accuracy of up to 95.89%, an AUC of up to 97.56%, a G-mean of up to 81.81%, and a F-measure of up to 81.45%. Adaboost was the top-performing algorithm across all performance measure, followed by random forest, logistic regression, rotation forest, kernel factory, naive Bayes, neural networks, and k-nearest neighbors. We found that all top three classifiers (i.e., adaboost, random forest, and logistic regression) performed equally in statistical terms. In terms of AUC, rotation forest also had equal statistical performance compared to adaboost. Our information-fusion sensitivity analysis indicates that the top predictors of romantic ties are mainly socio-demographic features, frequency and time-related features. In terms of communication types, we find that comments, likes, and tags on photos, albums and videos are the most important. Our findings provide important insights for academics and researchers that want to model social ties. Our research shows that the incorporation of disaggregated features is necessary to uncover the true effect on romantic ties, which is otherwise averaged out.

Chapter 4. The aim of this study is to (1) determine which social media platform (Facebook or Twitter) is most predictive of box office sales, (2) which algorithm performs best, and (3) which variables are important. To do so, we apply a holistic social media analytical approach consisting of two stages. The first stage compares the predictive performance of several models based on Facebook and Twitter data. To conduct a fair comparison between both social media platforms, we built two types of models for each platform. The first type includes marketer-generated content (MGC) and page popularity indicators (PPI), whereas the second type augments the first model with user-generated content (UGC). To make sure that our results are reliable we compare Facebook and Twitter over seven algorithms (i.e., regularized linear regression, k-nearest neighbors, decision trees, bagged trees, random forest, gradient boosting, and neural networks) using five times two-fold cross-validation. The second stage combines the information from

both platforms and algorithms using information-fusion sensitivity analysis. This stage determines which variables from which platform and from which data type are most important in driving predictive performance. The analysis shows that Facebook is more indicative of box office sales than Twitter in terms of RMSE, MAE, MAPE and R^2. Facebook models outperformed the Twitter models by at least 11% in RMSE, 13% in MAE, 14% in MAPE, and 43% in R^2. The analysis also shows that the performance of Facebook models is significantly better in a majority of the cases for both models with and without UGC. However, the addition of UGC next to MGC and UGC does not lead to a significant improvement in predictive performance. Next to comparing the predictive capability of Facebook and Twitter, we find that random forest is the top performing algorithm across all performance measures followed by bagged trees, gradient boosting, k-nearest neighbors, decision trees, linear regression and neural networks. The results of our information-fusion sensitivity analysis show that the number of Facebook page likes (a PPI variable) was the most important variable. Other important variables were the hype factor of Facebook comments and the number of positive comments. In general, we can say that UGC and PPI variables were the most important, in terms of word-of-mouth variables volume was more important than valence. Our findings provide important insights for researchers and practitioners that want to predict box office revenues. For practitioners our study can serve as a framework to determine which platform, algorithms, and variables to use when estimating box office revenues. For researchers our results offer both methodological and theoretical insights. From a methodological perspective, we provide insight into the most important variables and algorithms in box office sales predictions. From a theoretical perspective, we show which variables from which data type are most important and which marketing theories are most important.

Table 1.2: Overview of the contributions, main findings and practical implications of each study

Study	Contributions	Main findings	Practical implications
The Added Value Of Facebook Friends Data in Event Attendance Prediction	• Assess added value of friends data over and above user data in event attendance prediction • Determine top predictors of event attendance and uncover the underlying relationships	• Significant increase in AUC from 0.21%-points to 0.82%-points (in 3/5 algorithms) • Top drivers: timing of the event and (relative) number of friends attending the event	Facebook Inc., event promoters and practitioners can increase accuracy with including friends data
Evaluating the Importance of Different Communication Types in Romantic Tie Prediction on Social Media	• Disaggregate approach in modeling social ties on social media • Benchmark wide range of single classifiers and ensembles • Determine the most important predictors and the underlying relationships	• Accuracy up to 97.89%, AUC up to 97.56%, G-mean up to 81.81%, F-measure up to 81.45% • Top drivers: socio-demographic similarity and frequency and recency of commenting, liking, and tagging on photos, videos and statuses	Disaggregated variables should be included when predicting social ties to uncover the true relationship
Comparing the Ability of Twitter and Facebook Data to Predict Box Office Sales	• Determine which social media platform (Facebook or Twitter) is most indicative of box office sales • Determine which variables from which platform from which type are most important	• Facebook is significantly more predictive than Twitter. • User-generated content does not significantly improve predictive performance • Top drivers: Number Facebook page likes and the hype factor of Facebook comments	• Methodological: list of best platforms, algorithms, and variables • Theoretical: consumer engagement behavior and the awareness effect are most important theories

2

The Added Value Of Facebook Friends Data in Event Attendance Prediction[1]

Abstract

This paper seeks to assess the added value of a Facebook user's friends data in event attendance prediction over and above user data. For this purpose we gathered data of users that have liked an anonymous European soccer team on Facebook. In addition we obtained data from all their friends. In order to assess the added value of friends data we have built two models for five different algorithms (logistic regression, random forest, adaboost, neural networks and naive Bayes). The baseline model contained only user data and the augmented model contained both user and friends data. We employed five times two-fold cross-validation and the Wilcoxon signed rank test to validate our findings. The results suggest that the inclusion of friends data in our predictive model increases the area under the receiver operating characteristic curve (AUC). Out of five algorithms, the increase is significant for three algorithms, marginally significant for one algorithm, and not significant for one algorithm. The increase in AUC ranged from 0.21%-points to 0.82%-points. The analyses show that a top predictor is the number of friends that are attending the focal event. To the best of our knowledge this is the first study

[1]Based on: *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.*

that evaluates the added value of friends network data over and above user data in event attendance prediction on Facebook. These findings clearly indicate that including network data in event prediction models is a viable strategy for improving model performance.

2.1 Introduction

Facebook is a large-scale social media platform with 2.13 billion monthly active users and 1.4 billion daily active users [84] and has grown to the point of becoming an important channel for social contact [80, 156] and product promotion [23, 30]. Among other things, it enables businesses to schedule meetings and gatherings using a functionality called Facebook Events [82]. With Facebook Events promoters can manage event participants and notify participants' friends [82]. The downside of this functionality's popularity is that many companies are using it and hence there are a lot of co-occurring events [13]. In order to make a user's Facebook experience more enjoyable and to avoid information overload, Facebook predicts whether or not the user will attend the event. It logically follows then, that a very important task is to try and make those predictions as accurate as possible.

While there is a considerable body of research on event modeling in other fields and networks [56, 128, 163], little research has been done on Facebook Events specifically, despite the platforms' aforementioned size and success. A very common and important research question in event predictions pertains to the importance of specific sets of predictors. If a set of predictors does not improve predictive performance it should be removed from the model so as to prevent from slowing down the modeling process. In the case of Facebook data, a meaningful question is whether friends data should be included in the model. If a typical user has 300 friends, and we have 1000 users in our sample, including friends data would imply analyzing an additional 300,000 users. If these data do not improve the predictive model significantly, adding them would imply an unnecessary lag in the modeling process.

This paper seeks to fill this gap in literature by studying the added value of friends data over and above user data in event prediction on Facebook. We focus on predicting whether a soccer fan will declare to attend a given event or not. For this purpose we developed a Facebook application to extract a user's data along with a user's friends data. In total 5010 users and 1,102,573 friends authorized our application to collect their relevant data. To investigate the added value of friends data we build and compare two models. The first one includes only user data and the second one includes both user data and friends data . The difference in performance between both models yields the added value of friends data. If the performance increase is significant, friends data should be incorporated in future models. If not, it should be excluded for the sake of parsimony and execution

speed. Furthermore, we benchmark these two models for five state-of-the-art classification algorithms namely logistic regression, random forest, adaboost, neural networks and naive Bayes.

In the remainder of this article we first provide an overview of extant literature. Second, we provide details on the methodology. Third, we elaborate on our findings and their implications. Finally, we discuss limitations and avenues for future research.

2.2 Literature overview

The addition of social network information has proven to achieve good performance in several applications (other than event prediction). On Facebook, examples can be found in the field of activities [226], users [45], movies [186] and interests [105]. On Twitter, network information has proven to be useful in predicting user behavior [180] and tweet popularity [115, 203]. On other social network sites, including social relationship data has improved results in peer recommendations [146, 225]. Despite the importance of network data in social media prediction, literature on event attendance prediction remains scarce as discussed in the next paragraph.

Literature on event prediction can be classified according to the data that is used in the model. In this typology there are three classes: predictive models that are enriched with (1) user data [e.g., 163], (2) network data [e.g., 209], or (3) both user and network data [e.g., 116]. User data are defined as specific profile characteristics that represent the preferences of the user. Examples are the interests of the user [48], demographics [181] and past event-history [228]. Network data are defined as data that contain information about the user's social network. Examples are the number of peers that are attending the event [154], and event preferences of their friends [131].

Table 2.1 provides a literature review on event attendance prediction literature with a focus on data sources and platforms. It is clear that, to the best of our knowledge, our study is the only one that evaluates the added value of network data over and above user data on Facebook. Even more so, Table 2.1 indicates that the added value of network data has not been evaluated on other platforms. The study of Zhang et al. [228] is of special interest as it focuses on user and network data from Facebook, just as our study.

In their research, three large groups of event predictors and corresponding approaches are proposed. First, in a similarity-based approach (SBA) they use event profile data (e.g., topic, location) and user profile data (e.g., interests, activity history) to compute similarities. Second, in an approach that they call the relationship-based approach (RBA), they include network data such as whether or not friends will attend the event. Third, in their history-based approach (HBA) they

Table 2.1: Overview of events literature

Study	Case	Facebook data	User data	Network data	Added value network
Mynatt and Tullio [168]	Company meetings		X		
Horvitz et al. [116]	Company meetings		X	X	
Lovett et al. [154]	Company meetings			X	
Tullio and Mynatt [209]	Company meetings			X	
Daly and Geyer [56]	Company meetings		X	X	
Pessemier et al. [181]	Cultural activities	X	X		
Coppens et al. [48]	Cultural activities	X	X		
Lee [142]	Cultural activities		X	X	
Kayaalp et al. [128]	Concerts		X	X	
Minkov et al. [163]	Academic events		X		
Klamma et al. [131]	Academic events			X	
Zhang et al. [228]	Facebook events and Academic events	X	X	X	
Li [145]	Social event site		X	X	
Our study	Facebook events	X	X	X	X

add users' historic event attendances. The authors subsequently propose a hybrid approach (SRH), which is a combination of the three other approaches and data sources. Their research concludes that indeed the combination of all three data sources (SRH) yields the most precise and accurate results, followed by RBA, SBA and HBA.

Just as in the other studies in Table 2.1, Zhang et al. [228] do not assess the added value of network data over and above user data. They only investigate the difference in precision between the hybrid approach and the other methods. They have not made pairwise comparisons between the three different data sources by solely comparing the combined sources with the individual sources. Their results suggests that the SRH approach significantly outperforms the three other approaches. For the three other models, their study only states that they perform better than a random model, thereby neglecting to investigate whether the models are significantly different from one another. With this approach, they are also unable to detect whether the increase in performance is due to network data or not. Regarding these results, it is clear that their study does not incorporate a comprehensive assessment of the added value of friends data. Furthermore, their research doesn't disclose which variables should be included or not in order to make predictive models as efficient as possible. Such assessment is necessary because including friends data implies a certain computational cost. From that perspective, one could argue that including friends data is only reasonable if the results improve significantly.

To fill this gap in literature, this study focuses on one such pairwise compari-

son: it will assess the extra value of friends data over and above user profile data. By doing so, we can precisely isolate the impact of our network variables. To make the comparison we build two models, a first one -the baseline model- containing user predictors and a second one -the augmented model- with network predictors in addition to the user predictors[2]. Examples of user variables are the number of groups, posts, events and photos. Network variables are operationalized as the number and percentage of friends that are attending a certain event. Furthermore, we assess several algorithms to determine if the increase in prediction performance is consistent.

We have three hypotheses about why network variables might improve event recommendations. First, the theory of homophily [5, 160, 216], also called endogenous group formation [108], states that like-minded people group together and often share the same tastes and opinions [103, 200, 223]. Second, and closely related to homophily, is the idea of social influence [89] and selection [160]. The former states that persons tend to follow the decisions of their peers [52]. The latter states that people mostly select friends who are similar [87]. Third, network variables capture the concept of trust. Trust-based theories state that friends' actions will be more easily followed and hence be more accurate if they are sourced from a trustworthy connection or friend. This is especially important in the case of events because trust and acceptance are critical factors for actual event attendance [120, 143, 179]. In addition, Facebook friends are often real-life friends [80] and can therefore be deemed trustworthy ties.

Various studies confirm the result that adding social relationships increases the performance of predictive models in Facebook applications relating to romantic partnership [14] and link prediction [126]. Chang and Sun [42] also found evidence that network variables play an important role in location check-ins. Using Facebook data, they conclude that previous check-in behavior of the user and the check-ins of friends are the most relevant predictors of check-in behavior. Thus, if a friend is attending a Facebook Event, a user may be more inclined to attend as well. It is clear that from the theories of homophily, social influence and selection that the probability of adopting a given behavior rises when others in one's network have already adopted that behavior [4, 12, 52].

To summarize, we found strong indications in extant literature that the augmentation of user data with network data can improve the predictive power of our model. To the best of our knowledge this is the first study to look into this issue for the social network site Facebook. In the next section of this paper we will elaborate on our methodology.

[2]In the remainder of this paper, we will always refer to the model with only user data as the baseline model and to the model with user and friends data as the augmented model.

2.3 Methodology

2.3.1 Data

In order to extract data from Facebook, we made a Facebook application for a European soccer team. To stimulate usage of our application we offered a prize (i.e., a signed shirt of a famous soccer player) to the participants and asked three questions to determine the winner. The application was advertised several times on the Facebook fan page of the soccer team. In addition, the application was added to the main page tabs for added visibility. Application users were presented with an authorization box in which they had to give their permission before the data were gathered from their profile. The data were collected between May 7, 2014 and June 9, 2014. In total we collected 5,315 event observations (2,368 unique events) from 978 users. We also gathered data of 194,639 friends, which are used for the creation of network predictors. The response variable in our models is binary, with the value 1 if users indicated that they were attending and 0 otherwise. Of all our event observations attendance is 78.2%.

2.3.2 Predictors

The user-related variables are summarized in Table 2.2. The 'Like' variables in our study only relate to likes generated by users. 'Likes' are also only available for a page, band, app, or leisure activity. In the photo and video variables the affix 'created' points out that the photo or video was uploaded, or created and immediately uploaded with the Facebook app. Tags in photos refer to tags of the user himself/herself. The variable 'username' captures if a user has upgraded his/her username to an alphabetic identifier from the standard numeric identifier. Due to regulations on Facebook, we could only gather the twenty-five last albums, photos, videos, links, status updates, notes and check-ins. In order to alleviate this restraint, we calculated the frequency by time as to no users in our database reached this restriction. For the last seven days, we computed the frequency of status updates, photo and link uploads, for the last four months album uploads and check-ins were computed, and for the last year notes and video uploads were computed.

Table 2.2: Overview of predictors

Variable category	Variable
Demographic and identification variables	Age
	IND(gender)
	IND(email)
	IND(website)

Geographical variables	IND(hometown) IND(location)
Professional/ Educational variables	COUNT(languages) COUNT(work) COUNT(educations) IND(education type)
Social variables	COUNT(family) IND(sexual orientation) IND(relationship status) COUNT(OF 23 family relationship types) (e.g., aunt) COUNT(Friend connections) COUNT(Groups)
Personal variables	COUNT(favorite teams) COUNT(sports) COUNT(television) COUNT(music) COUNT(movies) COUNT(books) COUNT(activities) COUNT(inspirational people) COUNT(interests) COUNT(OF 10 television categories) (e.g., Show) COUNT(activity category) IND(OF 14 interests) (e.g., Design) IND(OF 23 sports) (e.g., Fitness) IND(bio) IND(quotes) IND(political) IND(religion)
General Facebook Account variables	Length Facebook membership Recency last update=REC(profile update created) MEAN(album privacy) Profile completeness=SUM(IND(37 profile variables)) IND(username) Time ratio=SDIET(all actions)/MIET(all actions)
Likes	COUNT(OF 188 like categories) (e.g., Musician/band) COUNT(likes) REC/MIET/SDIET(like created) COUNT(posts likes)
Statuses	COUNT(statuses) REC/MIET/SDIET(status updated)
Photos	COUNT(photos) REC/MIET/SDIET(photo created)
Videos	COUNT(videos) REC/MIET/SDIET(video created)
Albums	COUNT(albums) REC/MIET/SDIET(album created)

Events	COUNT(events) MIET/SDIET(event created) IND(event time == start day) IND(event time == end day) IND(event time == month) IND(event time == season) IND(event time == year) IND(event time == weekend) IND(event location) LENGTH(event time)
Links	COUNT(links) REC/MIET/SDIET(link created)
Check-ins	COUNT(check-ins) REC/MIET/SDIET(check-in created) IND(check in app)
Notes	COUNT(notes) REC/MIET/SDIET(note created)
Games	COUNT(games) REC/MIET/SDIET(game created)
Tags	REC/MIET/SDIET(photo user tags) COUNT(video user tags) COUNT(photo user tags) COUNT(check-in user tags) REC/MIET/SDIET(video user tags)
Comments made	REC/MIET/SDIET(photos/albums/statuses/links/check-ins comments) COUNT(photos/albums/statuses/links/check-ins comments)
Comments received	REC/MIET/SDIET(photos/albums/statuses/links/check-ins comments received) COUNT(photos/albums/statuses/links/check-ins comments received)

With IND: indicator, COUNT: frequency, REC: recency, MIET: mean inter-event time, SDIET: standard deviation inter-event time, LENGTH: length of the time interval. MIET is the mean time that passes between two subsequent events (e.g., album uploads). SDIET is defined as the standard deviation of the time between two subsequent events.

Within our user variables, we are particularly interested in event-related user variables. The majority of the user-event variables are calculated as time indicator variables (see Table 2.2 Section Events). These variables resolve to 1 if the event took place at a certain time and 0 otherwise. Applying this logic we computed dummies for the day of the week (for both start day and end day of the event), the weekend, the month, and the season. Other event variables such as the duration and location were also added. We denote that we didn't include dummies for the type of event, since our database mainly contains soccer events. Other popular events were related to parties and festivals. In total we calculated 540 user variables for our first model.

In order to create our second model, we augmented the first model with friends-

related variables. Next to our users we also gathered data from their friends (194,639). We computed five variables that are important for the event that we are predicting, namely the total and relative number of friends that are going to the focal event and the average number of total, soccer, and team events the user's friends attended.

2.3.3 Classification algorithms

In this section, we elaborate on the choice of our classification algorithms. In total, we use five single classifier and ensemble techniques: naive Bayes (NB), logistic regression (LR), neural networks (NN), random forest (RF), and adaboost (AB). Naive Bayes is the least complex algorithm because it only estimates the joint probability $p(x, y)$. In contrast logistic regression estimates the conditional probability $p(y \mid x)$ and this can result in better performance [170]. Neural networks are similar to logistic regression if the logistic activation function is employed but add additional complexity by incorporating a hidden layer. This increases flexibility and this can result in better performance. Random forest adds additional complexity by using an ensemble of trees. Trees are inherently nonlinear and incorporate interactions. Using many trees and combining them often improves performance. Finally adaptive boosting (adaboost) adds complexity by incorporating a weighting mechanism that focuses on incorrectly classified instances in the previous iteration. We will evaluate the added value of network variables for all these algorithms. This will allow us to draw conclusions across a range of complexity levels. In the following paragraphs we will provide more details about the different algorithms.

2.3.3.1 Naive Bayes

We use the original naive Bayes algorithm as a method for probabilistic classification. This method applies Bayes' Theorem to classify new observations and naively assumes conditional class independences [138]. Despite the fact that the conditional independence assumption is rarely satisfied, it achieves reasonable performance and low computation times [138]. Several authors have tried to overcome the problem of conditional dependency by introducing randomness such as random feature selection and bagging [139, 185]. The function *naiveBayes* was used from the *R*-package *e1071* [162]. Gaussian distributions were assumed for the numerical predictors.

2.3.3.2 Logistic regression

We use regularized logistic regression with the lasso approach to cope with overfitting. The lasso (least absolute shrinkage and selection operator) sets a bound on

the sum of the absolute values of the coefficients forcing the coefficients to shrink towards zero [122, p219]. In this regard, the value of the shrinkage parameter λ determines the amount of shrinkage. The higher the value of λ the smaller the coefficients will be. We use cross-validation to determine the optimal shrinkage parameter. The statistical *R*-package *glmnet* by Friedman et al. [92] is used to create our model. We set the parameter α to 1 to obtain the lasso approach and we set the $nlambda$ parameter to 100 (default) to compute the sequence of λ.

2.3.3.3 Neural networks

We use the feed-forward artificial neural network optimized by BFGS with one hidden layer. This approach is considered much more reliable, efficient and convenient than backpropagation and has proven to be sufficient in a variety of cases [70]. Before implementing the neural network, we rescale the numerical variables to $[-1, 1]$ [24]. The binary variables are disregarded and coded as $\{0, 1\}$. Scaling is necessary to avoid local optima and numerical problems and to ensure efficient training. The statistical *R*-package *nnet* is used to build the neural network [189]. The network weights are randomized at the start of the iterative procedure [190, p154]. This implies that the results change for subsequent neural networks, which mimics the development of the human brain [212]. We follow the recommendations of Ripley [190, p149] and set the *entropy* parameter to the maximum likelihood method. The *rang* parameter which manages the range of initial random weights was set to 0.5 (default). The parameters *abstol* and *rel* were also left at their default $1.0e^{-4}$ and $1.0e^{-8}$. Weight decay was used to avoid overfitting [70] and the maximum number of weights (*MaxNWts*) and maximum number of iterations (*maxit*) were set at a very large number (5000) in order to avoid early stopping. Finally a grid search was performed in order to determine the weight decay and the number of nodes in the hidden layer [70]. In accordance to Ripley [190, p163, p170] we sequenced over all combinations of $decay = \{0.001, 0.01, 0.1\}$ and $size = [1, ..., 20]$ to determine the optimal combination.

2.3.3.4 Random forest

Random forest combines bagging with random feature selection to build an ensemble of trees [35]. Each tree is grown on an independent bootstrap sample and at each node of each tree a randomly selected subset of features is evaluated [35]. To grow the ensemble all the trees are aggregated by means of majority voting [35]. As a result, random forest copes with the instability and the suboptimal performance of decision trees [74]. Two parameters have to be provided: the number of trees and the number of predictors randomly selected at each node of each tree [73, 140]. We follow the recommendation of Breiman [35] to use a large number of trees (500) and the square root of the total number of predictors as the num-

ber of predictors to be evaluated at each node. We use the statistical *R*-package *randomForest* provided by Liaw and Wiener [147].

2.3.3.5 Adaboost

The original adaboosting algorithm [91] sequentially reweights the training data [109, p337-340]. In each iteration the observations that were misclassified in the previous iteration are given more weight, whereas the correctly classified observations are given lower weight. Hence, instances that are hard to classify are given more importance in each iteration. The final model is a linear combination of all the previous models [109, p337-340]. We use stochastic boosting, one of the most recent boosting variants which introduces randomness as an integral part of the procedure [94]. Randomness is induced by making bootstrap samples in which the propensity of an observation being selected is proportional to the current weight [94]. There are three important parameters: the number of iterations, the number of terminal nodes in the base classifier, and the loss function. In accordance with Friedman [94] we determine the number of terminal nodes by setting the maximum depth of the trees to 3 and we set the number of iterations to 500. We use the exponential loss function to set the weights at each iteration. To fit our model we use the statistical *R*-package *ada* [54].

2.3.4 Performance evaluation

We use the area under the receiver operating characteristic curve (AUC or AUROC) to evaluate the performance of our classification models. AUC is argued to be an objective performance measure for classification problems by several authors [137]. The receiver operating characteristic curve (ROC) is a graphical representation of the sensitivity against one minus specificity for all possible cut-off values [106]. AUC is a more adequate measure of classifier performance than PCC (Percentage Correctly Classified) [16] whenever the cut-off value that will be used at model deployment is unknown, because AUC evaluates the entire range of cut-off values [17]. AUC is defined as follows:

$$AUC = \int_0^1 \frac{TP}{(TP+FN)} d\frac{FP}{(FP+TN)} = \int_0^1 \frac{TP}{P} d\frac{FP}{N} \qquad (2.1)$$

with TP: True Positives, FN: False Negatives, FP: False Positives, TN: True Negatives, P: Positives (event), N: Negatives (non-event).

Intuitively, AUC is the probability that a randomly chosen positive item is ranked higher than a randomly chosen negative item (i.e., the probability that a user who attends the focal event is ranked higher than someone who does not attend the event) [213]. AUC is restricted between the values of 0.5 and 1, where the former denotes that the model does not perform better than random and the

latter indicates a perfect prediction [106]. If there is a huge drop in AUC of the test set, this is a strong indication of overfitting.

2.3.5 Cross-validation

We use five times two-fold cross-validation (5x2cv) to make sure our results are not overly optimistic or pessimistic [1, 65]. 5x2cv starts by randomly dividing the sample in two parts where each part is used once as a training sample and once as a test sample. If the hyper-parameters of the algorithm require tuning, the training set was again split into two equal parts. After tuning, the original training set was used to build the final model. This process is repeated five times and results in 10 AUCs per model [65]. We take the median of the results to obtain the overall AUC of our models. As a measure of dispersion, the interquartile range (IQR) is used.

In order to test whether two models are significantly different from eachother we follow Demšar's [2006] suggestion to use the Wilcoxon signed rank test [218]. The Wilcoxon signed-rank test [218] is a non-parametric test that ranks the differences in performance of two models while ignoring the signs. Ranks are assigned from low to high absolute differences, and equal performances get the average rank. The ranks of both the positive and negative differences are summed and the minimum of those two is compared to a table of critical values. To be significant the smallest sum of ranks should be smaller than the critical value.

This test is considered safer than a parametric t-test because the assumptions of normality and homogeneity of variance [64] do not need to be met. However, when the assumptions of a t-test can be satisfied, the Wilcoxon signed rank test has less power than a paired t-test. When the sample size equals 10 verifying normality and homogeneity is problematic and thus the Wilcoxon signed rank test is preferred [64].

2.3.6 Variable importance evaluation

Because we are using a lot of predictors in our sample, it is important to know which variables have great predictive power [194]. One way to do so is by calculating the variable importances. In tree- based methods such as random forest we can evaluate the importance of our predictors by using the total decrease in node impurities from splitting on the variable, averaged over all trees. The Gini index is used as a measure of node impurity [36]. The importances are then averaged over the 10 folds by taking the median of the 5x2cv variables importances. We used the *importance* function in the *randomForest* package [147].

2.3.7 Partial dependence plots

Partial dependence plots allow one to graphically depict the relationship between an independent and a dependent variable, after eliminating the average effect of the other independent variables [93, 95]. This is analogous to multiple linear regression of y on all x_j, where the coefficient x_1 accounts for the effect of x_1 on y with the other variables kept constant. Partial dependence plots are mostly used on decision tree-based methods and allow one to gain insight in how classification variables relate to the most important predictors [95, 109, p369-370]. In order to create a partial dependence plot we follow the method described by Berk [28, p222].

For each value v in the range of a predictor *x* we create a novel data set where *x* only takes on that value. All the other variables are left untouched. Next, for each novel data set, we score all the instances using a Random Forest model that is built on the original data. Subsequently the mean of half the logit of the predictions is calculated yielding one single value for all instances called p. The final step in creating the partial dependence plot is plotting all the values v of x against their corresponding p. All partial dependence plots are five times twofold cross-validated using the *interpretR R*-package [21].

2.4 Discussion of results

2.4.1 Model performance

The cross-validated results are summarized in Figure 2.1 and Table 2.3. The main research question of this study was to assess if friends (i.e., network) data add value over and above user data in event prediction. We find that the inclusion of network variables results in an improvement of the AUC for all our classifiers. For the baseline model the AUC ranges from 65.21% to 79.56%, for the augmented model from 66.01% to 80.38%. The increase in AUC ranges from 0.21%-points to 0.82%-points. Figure 2.1 also reveals that adaboost (AB) is the top performing algorithm, followed by random forest(RF), logistic regression (LR), neural networks (NN) and naive Bayes (NB). However, for computational reasons one might prefer RF since it allows parallel execution whereas AB is sequential in nature.

The Wilcoxon tests (Table 2.3) indicate that the results are significantly different for three out of five classifiers. The results show a significant difference on the 1% significance level for RF, AB and NB and on the 10% significance level for LR. We found no significant difference for our NN classifier. Adding friends data results in a slight increase in interquartile range (IQR). Nevertheless the IQRs are low for all classifiers, indicating that all classifiers have stable results. The IQR also confirms that adaboost is the top performer because it has the smallest IQR. These findings confirm our hypothesis that Facebook friends data can significantly

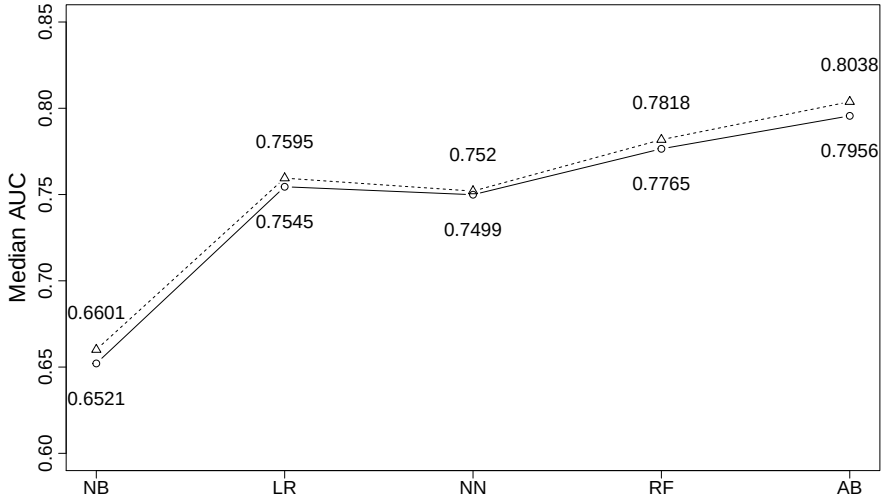


Figure 2.1: Cross-validated AUC. The solid line represents the baseline model, the dashed line the augmented model. NB = naive Bayes, LR = logistic regression. NN = neural networks. RF = random forest. AB = adaboost.

improve the predictions in event attendance prediction systems. It has to be noted though that for some classifiers results are not significant.

2.4.2 Predictors

In order to uncover what the main drivers of predictive performance are we first look at a scree plot of the predictors (Figure 2.2). In the scree plot, the 200 top predictors of the model with friends data are plotted against the median 5x2cv mean decrease in Gini in a descending order. It is clear from this plot that predictors with rank higher than twelve only add little to our predictions. Hence, we focus on the top twelve predictors in the rest of this discussion.

Table 2.3: Summary of cross-validated median AUC

	NB	LR	NN	RF	AB
Base model	0.6521	0.7545	0.7499	0.7765	0.7956
Augmented model	0.6601	0.7595	0.7520	0.7818	0.8038
Wilcoxon test	V = 0	V = 10	V = 21	V = 0	V = 0
	$p < 0.01$	$p < 0.10$	$p < 0.6$	$p < 0.01$	$p < 0.01$

Table 2.4: Summary of cross-validated median IQR

	NB	LR	NN	RF	AB
Base model	0.0039	0.0046	0.0086	0.0039	0.0035
Augmented model	0.0072	0.0160	0.0170	0.0057	0.0047

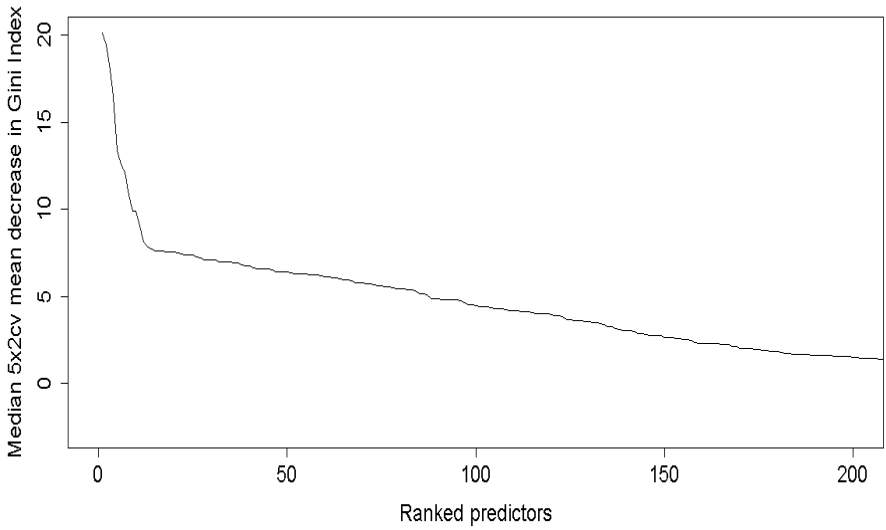


Figure 2.2: Scree plot of the 200 most important predictors

Table 2.5 contains the importance of the top twelve predictors and Figure 2.3 the partial dependence plots of selected variables. In Table 2.5 we observe that

Table 2.5: Median cross-validated variable importance

Rank	Variable name	Median decrease Gini
1	IND(event time == end day Mon)	20.107
2	IND(event time == start day Sun)	19.460
3	IND(event time == start month May)	18.023
4	COUNT(events)	16.447
5	IND(event time == weekend)	13.379
6	PERCENTAGE(friends event attending)	12.600
7	COUNT(friends event attending)	12.010
8	IND(event time == end day Sun)	10.814
9	IND(event time == start season Spring)	9.909
10	IND(event time == start day Sat)	9.878
11	IND (event time == start season Summer)	9.025
12	IND(event time == start month June)	8.146

most of the top predictors are related to the timing of the event and the friends variables. The most important predictor of event attendance is whether the event ends on a Monday. In Figure 2.3a we clearly observe a positive relationship between that predictor and event attendance. A plausible explanation can be found in the specific nature of our data. Major soccer events are mostly held on a Sunday. Hence event promoters on Facebook mostly set the ending of the event one day later (Monday). We also ran a plot (not shown) of whether the event starts on a Sunday and found the same positive relationship. Conversely, plots related to whether the event starts in the weekend depict a negative relationship with the probability of attending (not shown). This reinforces our explanation that important soccer games take place on Sunday (and their end time is always set to Monday on Facebook), minor soccer games are mostly played on other days in the weekend and receive less public attention. We denote that events with their end time on Monday, were not denoted as weekend events, this explains the negative relationship with the response variable. In Figure 2.3b, we note a positive relationship between whether the event starts in the month May and event attendance. The month May is also traditionally the play-off season in European soccer leagues. The same logic can be applied to explain the positive relationship between the Spring season and our response variable, since the month May lies in Spring season (not shown).

The results in Table 2.3 and Figure 2.1 already clearly indicated that friends data improve model performance. These results are substantiated in that network

predictors (the total and relative number of friends that indicated their attendance) are among the top ten predictors (sixth and seventh variable in Table 2.5). Looking at the partial dependence plots in Figure 2.3d and 2.3e, we first observe a positive and afterwards a negative effect, when more friends (more than 12 or 1.8%) are attending. The main reason for this relationship can be found in the News Feed Algorithm (NFA). Each time a friend interacts with something on Facebook, such as replying to an event invitation, a user gets notified in his or her News Feed. However, in order to avoid information overload Facebook limits the number of notifications for the same event. If a lot of friends are going, the NFA will stop propagating the message through the News Feed due to anti-spam regulations [81]. This implies that the probability of attending will first rise with every (close) additional friend that indicates attendance, and then decrease to normal once a given number of friends has been reached. Generally, these findings are partially different from the findings of Aral et al. [4] and Backstrom [12] who state that the adoption probability rises when friends already adopted. This partial difference is undoubtedly due to the many changes the NFA has undergone since these studies have been published. For example, Facebook recently increased their anti-spam regulations by hiding promotional posts in the user's News Feed [69]. In addition, Facebook users now have more control over what they see in their News feed [69].

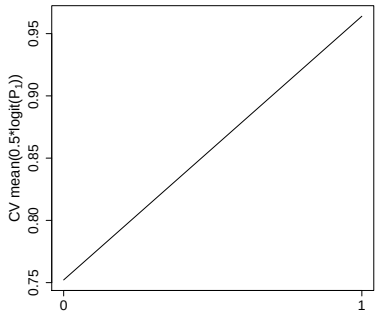
The total number of events the user attended (Figure 2.3c) is constant in the beginning and afterwards negatively related to our dependent variable. This implies that people will have an equal propensity of attending until they attend too many events. This is a plausible relationship. People don't have an unlimited amount of time to attend events. The more events the user attends, the less time he or she has to attend other events.

Finally, the predictors related to whether the event takes place in the Summer are negatively related with our dependent variable (see Figure 2.3f). Again, we refer to the specific nature of our data. In the Summer, the soccer season has ended and hence there are no important soccer events taking place. A diagnostic plot of whether the event is held in June and our response variable supports our hypothesis (not shown).

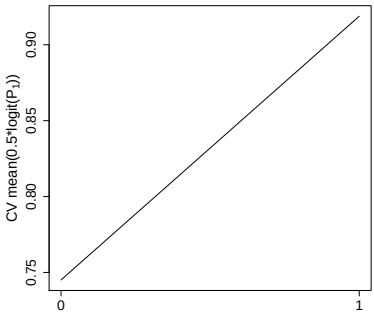
2.5 Conclusion and practical implications

In this study we set out to (1) evaluate the added value of a Facebook user's friends data over and above user data in event predictions and (2) gain more insight in the top predictors of event attendance.

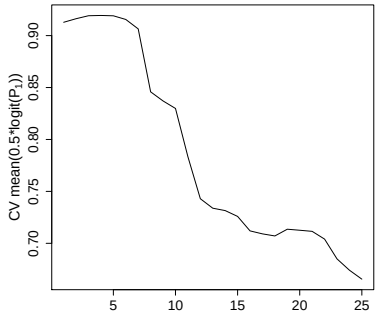
The results suggest that augmenting the data with network variables increases the AUC between 0.22%-points and 0.82%-points. This is in line with the conclusion of Benoit and Van den Poel [26] where the AUC also significantly rose with the inclusion of network effects. The top performing algorithm is adaboost,



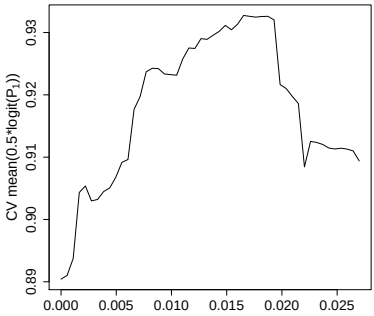
(a) Event end day: Monday



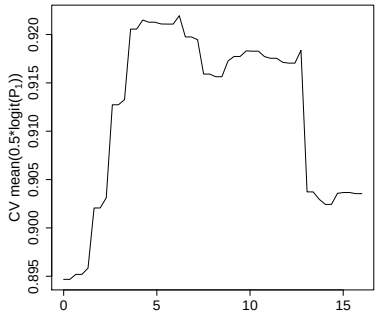
(b) Event start month: May



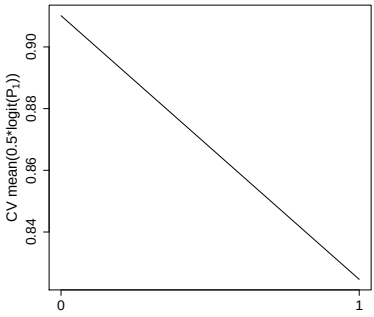
(c) Events attended in past



(d) Percentage friends attending event



(e) Friends attending event



(f) Event start season: Summer

Figure 2.3: Partial dependence plots

closely followed by random forest. This is similar to findings of [19], where adaboost and random forest came out as the best-in-class classifiers in a social media application. The top predictors are mainly related to the event time such as the start day and end day of the event. Network variables were also top predictors of event attendance. More specifically the absolute and relative number of friends that are attending the event are very important. We also provided a list of the top twelve predictors in Table 2.5 and partial dependence plots in Figure 2.3.

Our findings provide important insights for (1) Facebook Inc., (2) event promoters alike that want to increase the number of attendees, and (3) companies that want to build event prediction apps on Facebook. Facebook Inc. could incorporate our findings to adapt the News Feed Algorithm (NFA) for events. Recently, Facebook has finetuned the NFA algorithm to give more control to the user as to what he or she wishes to see and not to see in his or her News Feed [69]. Most of these updates are related to Facebook Pages and spam. Events however, are not specifically mentioned. A useful update could be to ask users to which extent they want to be informed about events, thereby giving them more control. Users seem to be positively influenced by a certain group of attendees. For example, users could indicate a threshold for the number of friends attending an event that controls when events appear and disappear in their News Feed.

Also event schedulers and promoters on Facebook can utilize our findings. Event organizers would benefit from more explicitly providing information about the attendees of an event. They can, for example, send invitations to friends of the attendees and include in the notification the number of friends that will attend.

Companies that want to create a Facebook app for event scheduling and promoting can also benefit from our results. We have proven that the inclusion of friends data significantly improves the accuracy of the prediction system. For example, when building an app that recommends events to a certain user, one could calculate which events are attended by his or her friends to generate more accurate predictions.

2.6 Limitations and future research

First, this study is limited because of selection effects. We extracted our data with a custom-built Facebook app via the Facebook Page of a European soccer team by offering a chance to win a prize. It might be the case that some users were not interested in this prize and hence were not willing to share their data. Another way of collecting data from Facebook is web-crawling as proposed by Lampe et al. [136] and Lewis et al. [144]. Nevertheless, web-crawling also suffers from the limitation that data cannot be extracted from private Facebook profiles. Generally, the collected data from web-crawling and a Facebook application largely overlap. Our approach is less intrusive since we ask permission from the user and

provide a 'rules and regulations' section in the app with our contact information. We also ensured the user that we anonymize all information and do not extract private messages. Finally, we also provided a disclaimer explaining the purpose of our academic research. Therefore, we believe that our approach is superior to web-crawling. Since we only limit our data to a subsample, our results do suffer from generalizability issues. However, regardless of this limitation our study is the first to investigate the added value of friends data in event attendance prediction. Hence, we consider this study a valuable contribution to literature. An avenue for future research can be to obtain a broader sample and more representative results.

A second limitation is that some of our predictors are limited in the number of values. Facebook only allows to extract the 25 most recent entries for specific variables. To mitigate this problem we computed the frequency of a specific time period as to no variable reaches this limit. The frequency of status updates, photo uploads and link uploads was calculated for the last 7 days, album uploads and check-ins for the last 4 months and video uploads and notes for the last year.

A third limitation is that we only include a limited number of friends variables in our analyses, mostly the ones that are related to the focal event. Following Zhang et al. [228], a possible avenue for future research could be to add more friends variables. We could investigate which type of predictors yields the biggest increase in model performance. This would help practitioners understand which elements in event attendance prediction systems make them as accurate and efficient as possible.

3

Evaluating the Importance of Different Communication Types in Romantic Tie Prediction on Social Media[1]

Abstract

The purpose of this paper is to evaluate which communication types on social media are most indicative for romantic tie prediction. In contrast to analyzing communication as a composite measure, we take a disaggregated approach by modeling separate measures for commenting, liking and tagging focused on an alter's status updates, photos, videos, check-ins, locations and links. To ensure that we have the best possible model we benchmark 8 classifiers using different data sampling techniques. The results indicate that we can predict romantic ties with very high accuracy. The top performing classification algorithm is adaboost with an accuracy of up to 97.89%, an AUC of up to 97.56%, a G-mean of up to 81.81%, and a F-measure of up to 81.45%. The top drivers of romantic ties were related to socio-demographic similarity and the frequency and recency of commenting, liking and tagging on photos, albums, videos and statuses. Previous research has largely focused on aggregate measures whereas this study focuses on disaggregate

[1]Based on: *Bogaert, M., Ballings, M., & Van den Poel, D. (2018). Evaluating the importance of different communication types in romantic tie prediction on social media. Annals of Operations Research, 263, 501 —527.*

measures. Therefore, to the best of our knowledge, this study is the first to provide such an extensive analysis of romantic tie prediction on social media.

3.1 Introduction

Literature on predicting social ties indicates that two types of variables are important: frequency-based variables [e.g., 40, 102, 157] and time-related variables [e.g., 75, 99, 148]. The importance of these variables holds across several social media applications such as mobile networks [224], e-mail communication [172], music platforms [25], Twitter [11], LinkedIn [221], Foursquare [178] and Facebook [126]. An example of a frequency variable in social media analytics is the number of comments from alter to ego [125]. An example of a time-related variable in social media analytics is the recency of communication [9]. Studies on both frequency-based and time-related variables [9, 100] indicate that time-related variables have greater predictive value than frequency variables.

While literature on tie prediction in social media has progressed a great deal it still suffers from one main limitation. Studies considering time-related communication variables in their predictive models mostly aggregate these variables over all interaction types and/or post types [9, 99, 100]. Nevertheless platforms such as Facebook offer a wide range of interaction types such as commenting, tagging and liking, and post types such as status updates, albums, photos, videos, check-ins, locations and links [144]. We call time-based variables, calculated for each interaction type and/or for each post type, disaggregated time-related communication variables. An example of such a feature is the recency of commenting on a photo. This is a time-based variable since we calculate the recency and it is disaggregated since we compute its value per interaction type (i.e., commenting) and per post type (i.e., a photo). There are several reasons why it is interesting to include disaggregated time-related variables. First, from a theoretical perspective it is key to know how tie prediction is impacted across the different interaction types. It could well be the case that time since last communication on Facebook is negatively related to tie presence, but this relationship might be stronger for certain interaction types. For example, time since last photo tag could be more indicative for predicting social ties than time since the last photo comment. The same reasoning holds for the different post types. For example, the time since last photo comment could have more predictive value than the time since last check-in comment. Hence, including disaggregated time-related predictors allows us to discover the true relationship with social ties. Second, from a predictive perspective, it could be that aggregate measures partially cancel out the effect of certain predictors. By including the variables separately we might improve our predictions.

To fill this gap in literature, we model social ties with disaggregated time-related and frequency variables. More specifically we will predict whether or not

ego and alter are in a romantic relationship. In other words we are focusing on romantic ties. We define disaggregated features as separate variables for comments, likes, and tags an ego has placed on alters' statuses, photos, albums, check-ins and location updates. An example of a disaggregated time-related variable is the elapsed time since the most recent comment of an ego on an alter's photo. An example of a disaggregated frequency variable is the number of comments of an ego on an alter's statuses. In addition to time-related and frequency variables related to communication, we also include similarity indicators: socio-demographic variables (e.g., indicator of common gender) and personal preferences (e.g., number of common interests, common events, and common groups). It is important to note that we will only focus on observable measures between ego and alter [9]. Hence, we do not include topological features of the user's social network such as the number of overlapping friends [59].

The remainder of this article is organized as follows. First, we review the literature on social tie prediction in social media. Second, we discuss the details of our methodology. Third, we conclude this study and discuss practical implications. Finally, we discuss the limitations and avenues for future research.

3.2 Related work

Predicting romantic ties is a subproblem of the more general social tie prediction problem, which in turn is a subproblem of the more general tie strength prediction problem. Therefore, in an attempt to highlight our contribution, we will discuss in this section the relevant literature on tie strength in social media. Literature on tie strength on social media can be categorized according to the type of communication predictors that are included in the model. In this study we define communication on Facebook as interacting with an alter's posts, which might or might not be directed to the ego. We only use publicly available information and hence exclude one-to-one communication. In this study, we focus on an ego's interaction activities such as commenting, liking and tagging focused on an alter's status updates, photos, videos, check-ins, locations and links. In this context we can characterize tie strength studies as follows: (1) studies including aggregated frequency indicators [e.g., 132], (2) studies with aggregated time-related features [e.g., 99], (3) studies that use disaggregated frequency predictors [e.g., 47], and finally (4) studies with disaggregated time-related features.

Aggregated frequency features are calculated as the total amount of interaction between ego and alter [8]. The higher the frequency, the stronger the relationship between alter and ego [11]. Aggregated time-related variables are measured as the elapsed time since the last communication on Facebook [9]. The shorter the recency (i.e., time since last communication), the better the relationship between two users [124].

Disaggregated frequency features are defined as the number of interactions for every interaction type, for every post type. An example of disaggregated features for frequency predictors are the number of likes on a photo or the number of comments on a status [125]. Disaggregated time-related features calculate the recency of all interaction types for every post type. Examples of disaggregated time-related features are the recency of a status comment and the recency of the last photo tag.

Table 3.1 provides an overview of the literature on tie strength prediction in social media with a focus on the type of features and the platform. It is clear from Table 3.1 that no study has modeled disaggregated time-related predictors.

Table 3.1: Overview of literature on tie strength in social media

Study	Case	Frequency		Time-related	
		Aggregated	Disaggregated	Aggregated	Disaggregated
Ogata et al. [172]	E-mail	X		X	
Kossinets and Watts [132]	E-mail	X			
Jeners et al. [124]	E-mail and Workspace platform	X		X	
Zhang and Dantu [227]	Mobile phone calls	X			
Xu et al. [224]	Mobile networks	X			
Choi et al. [47]	Online communication		X		
Baym and Ledbetter [25]	Music community	X			
Sheng et al. [199]	Micro-blog	X			
Arnaboldi et al. [8]	Twitter	X			
Baatarjav et al. [11]	Twitter	X			
Gilbert [99]	Twitter	X		X	
Liu et al. [150]	Twitter	X			
Kahanda and Neville [126]	Facebook	X			
Gilbert and Karahalios [100]	Facebook	X		X	
Xiang et al. [221]	Facebook and LinkedIn	X			
Arnaboldi et al. [7]	Facebook	X			
Zhao et al. [230]	Facebook	X			
Pappalardo et al. [178]	Facebook and Twitter and Foursquare	X			
Jones et al. [125]	Facebook		X		
Arnaboldi et al. [9]	Facebook	X		X	
Servia-Rodríguez et al. [195]	Facebook and Twitter	X			
Backstrom and Kleinberg [14]	Facebook	X			
Burke and Kraut [40]	Facebook		X		
Dunbar et al. [75]	Facebook and Twitter	X		X	
Trattner and Steurer [208]	My second life		X		
Wiese et al. [217]	Mobile networks	X		X	
Our study	Facebook		X		X

In order to address this gap in literature, we augment the existing prediction models with disaggregated time-related variables. We believe that this approach is better than using aggregated variables. Aggregated variables can hide the true effect of a certain feature on social ties. By using disaggregated features, we can clearly sort out which interaction types in combination with which post types are most important. This is confirmed by a recent study of Burke and Kraut [40] in which they study the evolution of ties over time. Their study reveals that the strength of the relationship increases over time with more personal communication such as comments and photos. Passive communication, on the other hand, such as liking does not influence the relationship between ego and alter. Based on the theory of social signaling [202], the authors argue that written communication demands more effort and brings people closer together. These findings are in line with earlier research on Facebook communication [61, 135] in that more vivid and interactive Facebook posts (i.e., status updates and photos) receive more interaction such as likes and comments. This implies that more entertaining posts will have a smaller recency of interaction than less entertaining posts, and that including separate variables, as opposed to aggregate variables, are likely to capture relationships that would otherwise be averaged out.

Based on the research discussed above we believe that disaggregated (time-related) variables will play a substantial role in determining romantic ties. It is important to note that we do not incorporate measures related to the whole social network (e.g., embeddedness or network centrality) in our analysis [14]. Whereas social network analysis focuses on the features of the whole network and the relationships between alters [59], our study analyzes the individual characteristics of the relationship between ego and alter [114, 191].

In summary, we found indications in extant literature that the inclusion of disaggregated time-related variables can improve prediction models of romantic relationships. To the best of our knowledge, we are the first to include time-related features for each interaction type, next to the existing disaggregated frequency predictors. In the next section, we will explain our methodology in more detail.

3.3 Methodology

3.3.1 Data

Above all other social media platforms, Facebook is considered the most important [129]. In comparison to other social media platforms (e.g., Twitter, LinkedIn, Instagram), significantly more information about the interactions between ego and alter can be gathered [136]. Hence, we chose Facebook as our platform of interest. In order to collect our data, we developed a customized Facebook application. The link to the Facebook application was promoted on Facebook several times. When

users opened the application they were first presented with an authorization box, which asked permission from the users to extract their data. In addition, a full and comprehensive list of the collected data was presented to the users along with an explanation that we would only use their data for academic purposes. If the users agreed to these terms, the data were collected from their profile. The data were gathered between May 7, 2014 and June 9, 2014. In total we collected data from 5006 users.

3.3.2 Variables

Our dependent measure is whether or not an alter is an ego's significant other. On Facebook, users can indicate whether or not they are in a relationship, engaged, married or single. Two users that have indicated to be in relationship, engaged or married with each other, are seen as significant others. In total, 816 users indicated that they had a significant other.

Table 3.2 summarizes the predictors used in our model. In Table 3.2 IND stands for indicator, COUNT stands for frequency, and REC stands for recency. REC calculates the time since the last communication. Next to disaggregated frequency and time-related features, we also included similarity variables (socio-demographic and personal preference variables) summing up to a total of 49 predictors. Socio-demographic similarity variables were computed as binary variables which resolved to 1 if ego and alter had an overlapping socio-demo characteristic (e.g., location), and 0 otherwise. The personal preference similarity variables were defined as the total number of items ego and alter had in common. An example is the total number of events both ego and alter attended. The disaggregated frequency variables included in our model refer to the total number of times an ego has tagged, commented on, or liked an alter's album, check-in, link, photo, post, status or video. Tagging refers to the total number of times an ego has been tagged in an alter's check-ins, locations, photos or videos. For the disaggregated time-related variables, we computed recency or time since last communication measured per interaction and post type. Recency is considered to be a good predictor for modeling social ties since it represents the intensity and the intimacy of the relationship [9, 100].

3.3.3 Data sampling

A problem in social tie prediction on social media is that a user only interacts regularly with a limited number of friends. This subgroup in an ego's social network is most often referred to as the active network [75, 76]. In the case of a binary response variable (e.g., predicting romantic ties), this is translated into a high class imbalance. For example, in our data only 0.07% (814 out of 1,102,573) of all collected relationships were identified as each other's significant other. Hence, it is

Table 3.2: Overview of features

Variable category	Variable
Socio-demographics similarity variables	IND (common gender/location)
Personal preference similarity variables	COUNT (common books/education/events/groups/interests/ likes/movies/music/sport/television/videos/work)
Disaggregated Frequency variables	COUNT (comments on album/check-in/link/ photo/post/status/video) COUNT (likes on album/check-in/link/ photo/post/status/video) COUNT (tags on check-in/location/ photo/video)
Disaggregated Time-related variables	REC (comments on albums/check-ins/ links/photos/posts/statuses/videos) REC (tags on photo/videos)

necessary to create a new sample, which represents the active network of our users. First, since our dependent measure focuses on romantic relationships, we only included users who indicated on Facebook that they were in a relationship, engaged or married. Next, we computed the interaction level for our non-romantic relationships over all disaggregated frequency variables. The interaction level is determined by taking the mean across all disaggregated frequency variables for each non-romantic relationship. By doing so, we have an overall measure of interaction for each non-romantic relationship between ego and alter. Table 3.3 provides the summary statistics for the interaction level of all non-romantic ties. We only selected non-romantic friendships with an interaction level higher than 3.945 (3^{rd} quartile). Hence we solely included non-romantic relationships with a high level of interaction. Finally, we imposed that the number of men and women had to be equal in the non-romantic relationships because we wanted to avoid that an oversampling of men or women would skew the results in favor of one gender. Hence, our final sample consists of 6720 relationships of users and friends, of which 814 (12.11%) indicated to be each other's significant other and 5906 (87.89%) did not. It is important to note that, because we filtered out alters that were very unlikely to be a user's significant other (for the sake of obtaining class balance), we made it more difficult for the model to make overall correct predictions. By removing the easy to classify cases we effectively made the task harder for our model. Hence it is safe to say that our model evaluations are conservative.

In order to cope with the remaining class imbalance problem (i.e., the underrepresentation of the romantic ties in our data set), we use several data sampling techniques: random undersampling (RUS), random oversampling (ROS) and synthetic minority oversampling technique (SMOTE) [67]. Random undersampling randomly drops instances from the majority class, without removing examples

Table 3.3: Summary statistics of interaction level of all non-romantic friendships

Min.	1st Quartile	Mean	Median	3rd Quartile	Max.
0.511	2.536	3.097	3.285	3.945	9.359

from the minority class [39]. In our case we remove as many instances from the majority class until the distribution across both classes is equal. Random oversampling, on its part, randomly replicates instances from the minority class until a certain distribution is reached (in our case a 50/50 distribution) [110]. Finally, SMOTE combines oversampling and undersampling by creating synthetic examples of the minority class and removing a certain percentage of the majority class [44]. Synthetic samples are created by (1) randomly selecting one of the k nearest neighbors of the selected instance, (2) taking the difference of the instance and its nearest neighbor and multiplying it with a certain number between 0 and 1 and (3) adding this difference to the selected instance [44]. We set the percentage of synthetic oversampling to 200% and adapted the percentage of undersampling in such a way that an equal distribution was achieved.

3.3.4 Prediction algorithms

In this section, we elaborate on the different modeling techniques used to estimate romantic ties. These techniques can be divided into single classifier techniques and ensemble methods. In total we applied four single classifier techniques: k-nearest neighbors (KNN), naive Bayes (NB), logistic regression (LR) and neural networks (NN). Next to algorithms that generate single classifiers, we also employed four ensemble algorithms: random forest (RF), adaboost (AB), kernel factory (KF), and rotation forest (RoF). Ensemble methods combine several single classifiers by means of averaging their predictions [66]. In that way, ensemble methods cope with the statistical, computational and representational problem of single classifiers [66]. On the one hand, ensemble classifiers reduce the variance of single classifiers by solving the statistical and computational problem. On the other hand, ensemble methods induce a reduction in bias by tackling the representation problem.

3.3.4.1 K-nearest neighbors

For the implementation of k-nearest neighbors (KNN), we chose the k-d tree algorithm [27]. In KNN, the parameter k represents the number of nearest neighbors to select when classifying new data. The prediction of the new data will be the proportion of the positive instances of the k samples. As a result, it is important to

determine the optimal value for k. Hence, we cross-validated our parameter k by sequencing over all values of $k = \{1, 2, 3, ..., 149, 150\}$. We used the *R*-package *FNN* to implement KNN [29].

3.3.4.2 Naive Bayes

As a method for probabilistic classification, we apply the generative naive Bayes algorithm which estimates a model based on the joint probability $p(x, y)$ [170]. In order to make predictions the classifier uses Bayes' theorem to calculate the conditional probability $p(y|x)$ by assuming conditional class independence [138, 170]. The naive Bayes classifier is a popular method because it yields good performance for low computation times [138]. However, the assumption of conditional class independence rarely holds. For the implementation of the naive Bayes classifier, we used the $naiveBayes$ function from the *R*-package *e1071* [162]. To calculate Bayes' theorem for numerical predictors we assumed Gaussian distributions.

3.3.4.3 Logistic regression

To avoid overfitting we used logistic regression with lasso regularization. The least absolute shrinkage operator restricts the sum of the absolute values of the coefficients with the shrinkage parameter λ [122, p.219]. In this regard, the parameter λ determine to which amount the parameters are shrunk towards zero. The higher the value of λ, the higher the shrinkage and the closer the coefficients will be to zero. The lasso approach can be seen as a form of variable selection which induces a small increase in bias in exchange for a decrease in variance [122, p.219]. We cross-validated the parameter λ in order to determine the optimal level of shrinkage. We used the *glmnet* package in *R* to estimate our model [92]. In order to implement the lasso approach we set the parameter α to 1 and the *nlambda* to its default which iterates over 100 λ values to calculate the sequence of λ.

3.3.4.4 Neural networks

To implement the neural network classifier, we used the feed-forward propagation algorithm optimized by BFGS with one hidden layer. This method is considered as more robust, efficient and easier to implement than the backpropagation algorithm. As a consequence the algorithm can be used in a large variety of cases [70]. In order to overcome computational issues and increase training efficiency we rescale the numerical predictors to $[-1, 1]$, and the binary variables are transformed to $\{-1, 1\}$. Scaling is also necessary to avoid local optima. Not scaling the data would cause the hyperplanes to only separate a limited amount of data, which increases the probability of local optima. We use the *R*-package *nnet* to model our neural network [189]. We follow the recommendations of Ripley [190] to imple-

ment our neural network. We start the iterative procedure by randomizing the network weights, such that the subsequent runs can produce different outcomes [190, p.154]. We set the *rang* parameter to its default of 0.5, which determines the range of the random weights in the beginning. Following Spackman [201], we used the maximum likelihood algorithm for the *entropy* parameter. We set the parameters *abstol* and *reltol* to their corresponding default values of $1.0e^{-4}$ and $1.0e^{-8}$. To keep the algorithm from early stopping, the maximum number of iterations (*maxit*) and the maximum number of random weights (*MaxNWts*) were assigned a very high value (5000). To avoid overfitting, we applied weight decay [70]. The weight decay parameter and number of hidden nodes were optimized by performing a grid search [70]. We iterated over all values of $decay = \{0.001, 0.01, 0.1\}$ and $size = \{1, 2, 3, ..., 20\}$ [190, p.170] and selected the optimal values.

3.3.4.5 Random forest

Random forest builds an ensemble of trees by combining bagging and random feature selection [35]. Bagging stands for bootstrap aggregating and means that each tree is grown on a bootstrap sample [34]. In addition, random feature selection requires that the best split at each node is determined by a random subset of predictors [35]. The final ensemble of trees is built by aggregating the trees through majority voting [35]. As a consequence, random forest adds an extra level of diversity to the modeling process and deals with the high instability, correlation and variance of decision trees [122]. The algorithm is easily implemented since only two parameters have to be provided (i.e., the number of trees to grow and the number of predictors to randomly select at each node split) [19, 23]. We follow the recommendations of Breiman [35] and set the number of trees to 500 and the number of predictors to be evaluated at each split to the square root of the total number of predictors. We use the statistical *R*-package *randomForest* to implement the algorithm [147].

3.3.4.6 Adaboost

We use the stochastic boosting algorithm, one of the most recent variations of adaboost [94]. As the original adaboost algorithm, stochastic boosting uses weighting to sequentially create new instances of the training data [91]. In each sequence misclassified instances are assigned more weight for the next sequence, whereas correctly classified cases are given a lower weight. The final model is a linear combination of all the models created in the previous sequences [122, p.337-340]. In contrast to the original adaboost approach, stochastic boosting improves the procedure by incorporating randomness [94]. More specifically, stochastic boosting creates bootstrap samples with the probability of an observation being selected equal to the weight of that observation [94]. We follow the suggestions of Fried-

man [94] and set the maximum depth of the trees to 3 for the number of terminal nodes and allow 500 iterations. We choose the exponential loss function to determine the weight of each instance at each iteration. We created our stochastic boosting model with the *R*-package *ada* [54].

3.3.4.7 Kernel factory

Kernel factory is an ensemble method for kernel machines [18]. First, the training data is randomly divided into mutually exclusive samples. Next, the burn method is used to automatically select the best kernel function and transform each sample into a kernel matrix *K*. Each kernel matrix *K* is then used as input training data for a random forest model, which generates a number of predictions equal to the number of samples. The final predictions are generated by taking the weighted average of the random forest predictions. Furthermore, a genetic algorithm is applied to find the optimal weights. The recommended values for the column and row partitions are respectively 1 and $log_{10}(N)$. We implemented kernel factory using the *R*-package *kernelFactory* [20].

3.3.4.8 Rotation forest

Rotation forest is based on the technique of feature extraction [192]. In order to create the training data, the predictors are divided into *K* subsets [58]. Subsequently, principal component analysis (PCA) is used on each of the *K* subset. Next for each subset a decision tree is built including all the principal components. The final prediction is created by taking the average of each of the *K* predictions. There are two important parameters for rotation forest: the number of variables to select for each subset (*K*) and the number of base classifiers (*L*). We follow the recommendations of Rodriguez et al. [192] and set the number of variables to 3 and the number of base classifiers to 10. In order to implement rotation forest we used the *R*-package *rotationForest* [22].

3.3.5 Performance evaluation

In order to assess the performance of our prediction model, we use the following common measures: accuracy, G-mean, F-measure and the area under the receiver operating characteristic curve (AUC or AUROC) [67]. Consider a binary classification problem, where $\{P_o,N_o\}$ stands for the observed positive and negative instances and $\{P_p,N_p\}$ for the predicted positive and negative instances. We can then easily represent the classification performance by means of a confusion matrix (or contingency table) given in Figure 3.1 [110]. In Figure 3.1 *TP* stands for true positives, *FP* for false positives, *FN* for false negatives, and *TN* for true negatives. In our case the positive class is defined as being an alter's significant other.

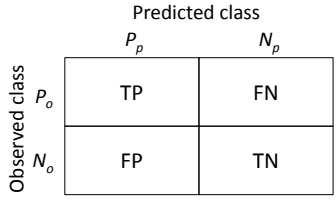


Figure 3.1: Confusion matrix

Accuracy, or the percentage correctly classified, is one the most used performance measures and can be defined as [110]:

$$Accuracy = \frac{TP + TN}{P_o + N_o}. \tag{3.1}$$

Since accuracy is sensitive to the distribution of the data, it is not fit as a performance measure for imbalanced settings [110]. To overcome this problem precision, recall, F-measure and G-mean can be used [110]. These measures can be defined as follows [67]:

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{P_p}, \tag{3.2}$$

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{P_o}, \tag{3.3}$$

$$F - measure = \frac{(1 + \beta)^2 \cdot Recall \cdot Precision}{\beta^2 \cdot Recall + Precision}, \tag{3.4}$$

where β determines the relative weight given to precision in comparison to recall (we set $\beta = 1$), and

$$G - mean = \sqrt{Recall \cdot Precision} = \sqrt{(\frac{TP}{P_o})(\frac{TP}{P_p})}. \tag{3.5}$$

We note that we only allow for probabilistic output since a ranking is required of people who are most likely to be one's significant other. For accuracy, F-measure and G-mean a specific threshold (i.e., operating condition) needs to be chosen that classifies an instance in the positive or negative class [113]. This threshold has to be proportional to the percentage of people that we want to target [19]. Since we are not interested in recommending a large proportion of an ego's network as a significant other, we will select the top 10% users most likely to be a significant other. Hence we set our threshold to the value that results in a proportion of 10%.

However, often the future operating conditions in which a classifier has to be deployed are unknown. A solution in that case is to use aggregate measures which aggregate over a distribution of all possible cutoff values [113]. The most-well known example of portmanteau measures is the area under the receiver operating characteristic curve (AUC or AUROC). AUC is considered to be a more appropriate and objective performance measure when the cut-off value is unknown at the time of model evaluation. The receiver operating characteristic curve (ROC) graphically depicts the relationship between the sensitivity and one minus the specificity for the entire range of cut-off values [106, 206]. Intuitively, AUC is the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance [50]. AUC ranges between the values of 0.5 and 1. The former indicates that the model performance is not better than random, the latter implies that the prediction is perfect [106]. The AUC is defined as follows:

$$AUC = \int_0^1 \frac{TP}{(TP+FN)} d\frac{FP}{(FP+TN)} = \int_0^1 \frac{TP}{P} d\frac{FP}{N}. \tag{3.6}$$

3.3.6 Cross-validation

In order to ensure the robustness of our results, we use five times two-fold cross-validation (5x2cv)[1]. First, 5x2cv randomly splits the data into two folds. Next, each fold is used once as a training and once as a test set. If tuning of the hyper-parameters is required, the training data is split again into two equal parts, namely a training and a validation set. After parameter tuning on the validation set, the initial training set was used to build the model and the test set to evaluate performance. Finally, the procedure is iterated five times until there are ten different performance measures [64]. As a measure of overall performance, we report the median accuracy, G-mean, F-measure and AUC of the ten different models. In addition, we also include the interquartile range (IQR) as a measure of dispersion.

In order to test for significant differences between the different modeling techniques, we use the non-parametric Friedman test [96] with Bonferroni-Dunn post hoc test [64, 77]. Within each fold the classifiers are ranked. Rank 1 is assigned to the best algorithm, rank 2 is assigned to the second best algorithm, ..., and the rank equal to the number of algorithms (in our case 8) is assigned to the worst performing algorithm. In case of ties the algorithms receive the average rank. The Friedman statistic is defined as [64]:

$$\chi_F^2 = \frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right] \tag{3.7}$$

with N the number of folds (in our case 10), k the number of different classifiers

(in our case 8), and $R_j = \sum_i^N r_i^j$. R_j is the average rank of the j^{th} algorithm and r_i^j stands for the rank of the j^{th} of k algorithms on the i^{th} of the N folds.

The null hypothesis of the Friedman test states that there is no difference between the different algorithms [96]. If the null hypothesis of the Friedman test is rejected, we calculate the Bonferroni-Dunn post hoc test to compare the different classifiers to a control classifier [77]. The Bonferroni-Dunn test is preferred over the Nemenyi post hoc test [169] since it has greater power when all classifiers are compared to a control classifier and not between each other [64]. Two classifiers are defined as statistically different when their average ranks differ by at least the critical difference (CD):

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} = 2.690 \sqrt{\frac{8(8+1)}{6.10}} = 2.9469 \tag{3.8}$$

with q_α the critical value for a given significance level α (in our case 0.05), k the number of different algorithms (in our case 8) and N the number of folds (in our case 10). In our study two classifiers are statistically different if their mean ranks differ by more than 2.9469.

3.3.7 Information-fusion sensitivity analysis

Since there is no single best method that works for every data set in predictive modeling, researchers often combine the results of different algorithms to obtain more precise results [66]. In that regard information fusion is defined as the process of fusing the information received from multiple prediction models [196]. Combining prediction models leads to more useful information and reduces the bias and uncertainty related to individual prediction models [175]. Let y represent our dependent variable and **x** our independent variables $\{x_1, x_2, ..., x_n\}$, then a single predictive model *i* can be written as:

$$\hat{y}_i = f_i(x_1, x_2, ..., x_n) = f_i(\mathbf{x}). \tag{3.9}$$

We note that f can take on any functional form. If k denotes the number of predictions models (in our case 8), our information-fusion model can be represented as follows [175]:

$$\hat{y}_{fusion} = \Psi\left(\hat{y}_1, \hat{y}_2, ..., \hat{y}_k\right) = \Psi\left(f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_k(\mathbf{x})\right), \tag{3.10}$$

where Ψ is the fusion operation. If Ψ is a linear combination function then Eq. 3.10 can be rewritten as follows:

$$\hat{y}_{fusion} = \sum_{i=1}^{k} \alpha_i f_i(x) = \alpha_1 f_1(\mathbf{x}) + \alpha_2 f_2(\mathbf{x}) + ... + \alpha_k f_k(\mathbf{x}), \tag{3.11}$$

where α_i is the weighting coefficient of each individual prediction model $f_i(\mathbf{x})$ and $\sum_{i=1}^{k} \alpha_i = 1$. The values of α are proportional to the predictive performance of each prediction model $f_i(x)$. Hence the better the predictive performance of the individual classifier, the higher their respective weight in the fusion function Ψ [175]. This implies that information extracted from well performing algorithms will receive higher importance than poorly performing algorithms. In our case, we will use median 5x2cv AUCs of the different algorithms as our weighting coefficients α.

When working with a lot of different predictors it is utterly important to know the rank order of the importance of the predictors [196]. In data mining variable importance measures are often seen as a form of sensitivity analysis, since they indicate how sensitive the model performance is to permuting on a certain predictor variable [175]. The more sensitive the predictive model is to a certain predictor, the higher its variable importance will be. Several measures of variable importance can be derived such as the mean decrease in Gini coefficient and the mean decrease in accuracy [23, 35]. However these measures tend to have sub-optimal performance when confronted with imbalanced data sets [123]. In order to overcome this problem Janitza et al. [123] have suggested an AUC-based permutation variable importance measure that is more robust than the traditional variable importance measure in an imbalanced setting. This measure is very similar to the traditional error-based variable importance but instead of using the error rate, we now use the AUC to measure the predictive performance. We can then reformulate Eq. 3.11 to an information-fusion based sensitivity measure of the predictor n with k prediction models (i.e., mean decrease in AUC after permuting on the variable n) as:

$$V_{fusion,n} = \sum_{i=1}^{k} \alpha_i V_{i,n} = \alpha_1 V_{1,n} + \alpha_2 V_{2,n} ... + \alpha_k V_{k,n}, \tag{3.12}$$

where $V_{i,n}$ stands for the variable importance measure of predictor n in prediction model i. The values of α are the same as in Eq. 3.11: the median 5x2cv AUCs of the different algorithms.

3.4 Results

3.4.1 Model performance

In order to determine which data sampling technique is superior we count the absolute (and relative) number of algorithms on which the data sampling technique is the overall winner for each performance measure [64]. The results of the total (and relative) number of wins across all 8 algorithms are summarized in Table 3.4.

Table 3.4: Absolute (and relative) number of wins across 8 algorithms based on accuracy, G-mean, F-measure and AUC for each sampling technique

	Accuracy	AUC	G-mean	F-mean
RUS	0 (0.00)	2 (0.25)	0 (0.00)	0 (0.00)
SMOTE	2 (0.25)	1 (0.125)	2 (0.25)	2 (0.25)
ROS	6 (0.75)	5 (0.575)	6 (0.75)	6 (0.75)

For the sake of completeness we provide the results of the median 5x2cv results of accuracy, G-mean, F-measure and AUC for each algorithm in 3.8. For example, in the case of accuracy we see that for 6 out of the 8 algorithms (75%) ROS outperforms RUS and SMOTE, for the remaining 2 algorithms SMOTE outperforms ROS and RUS. From Table 3.4 it is then clear that ROS outperforms RUS and SMOTE for the other remaining performance measure with respectively 5 wins for AUC, 6 for F-measure and 6 for G-mean. Hence, in the following sections our analysis will be based upon the 5x2cv results of ROS.

Figure 3.2, 3.3, 3.4 and 3.5 clearly indicate that we can predict romantic ties with high predictive performance when disaggregated time related and frequency variables are included in our model. The accuracy ranges from 95.89% to 86.64%. The AUC ranges from 97.55% to 75.49%. The F-measure, on its part, ranges from 81.45% to 50.4% and the G-mean from 81.81% to 50.8%. These AUC and accuracy values are in line with previous work on social tie prediction. For example, Jones et al. [125] build several classifiers (logistic regression, support vector machines, random forest) with disaggregated frequency, and socio- demographic and personal preference similarity variables as independent measures. In accordance with our study they use a binary dependent measure (i.e., is the alter a close friend of ego or not). They report a 5-fold cross validated AUC ranging from 73% to 92% and accuracy from 69% to 86%. Based on Figure 3.2, 3.3, 3.4 and 3.5 adaboost (AB) comes out as the top performing classifier, followed by random forest (RF) and logistic regression (LR), rotation forest (RoF), kernel factory (KF), naive bayes (NB), neural networks (NN) and k-nearest neighbors (KNN). Our results also confirm the findings of Bogaert et al. [32] and Ballings and Van den Poel [19] where adaboost also achieves the highest performance in social media studies in the case of event attendance prediction and Facebook usage increase.

Table 3.5 provides the average ranks of the classifiers together with the results of the Friedman test. The Friedman test points out that we can reject the null hypothesis for accuracy ($\chi_2(7) = 66.93, p < 0.001$), AUC ($\chi_2(7) = 67.57, p < 0.001$), G-mean ($\chi_2(7) = 67.1, p < 0.001$) and the F-measure ($\chi_2(7) = 67.1, p < 0.001$), so we need to perform the Bonferroni-Dunn post hoc test to find out which algorithms are significantly different from the top performing algorithm (i.e., adaboost). Based on the average ranks of the accuracy, AUC, G-mean and F-measure,

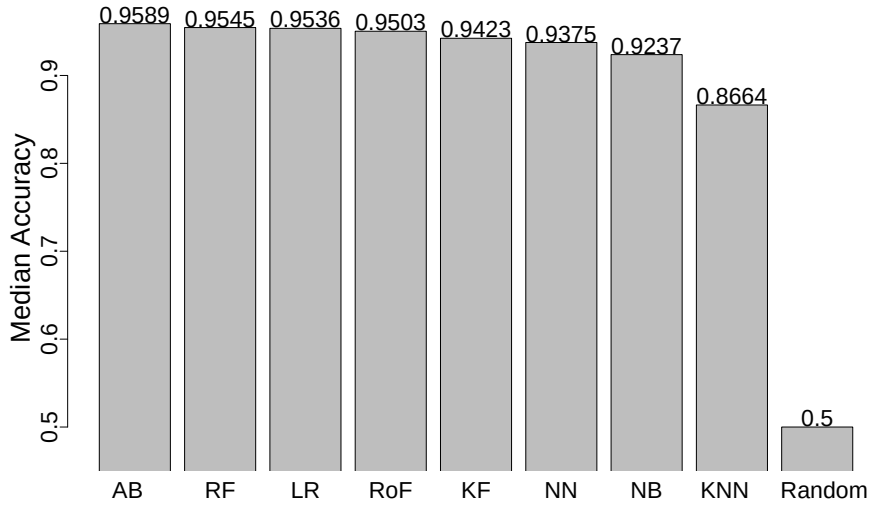


Figure 3.2: 5×2*cv median accuracy*

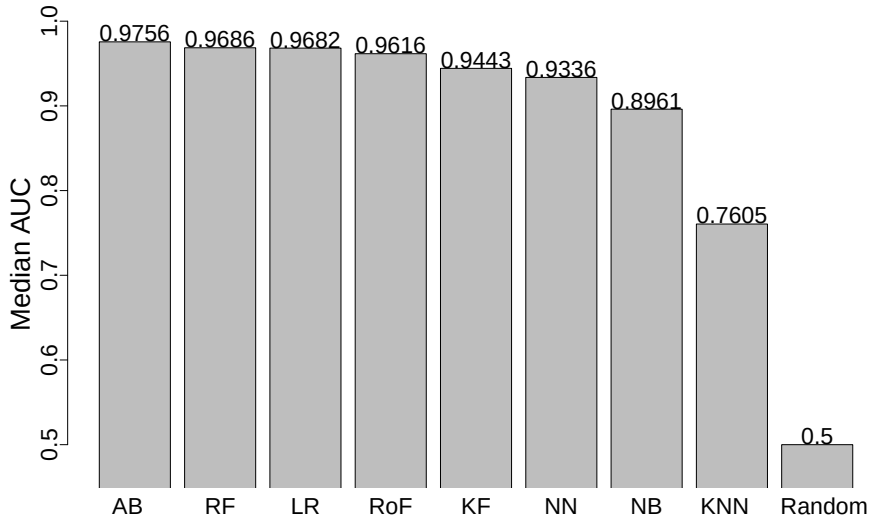


Figure 3.3: 5×2*cv median AUC*

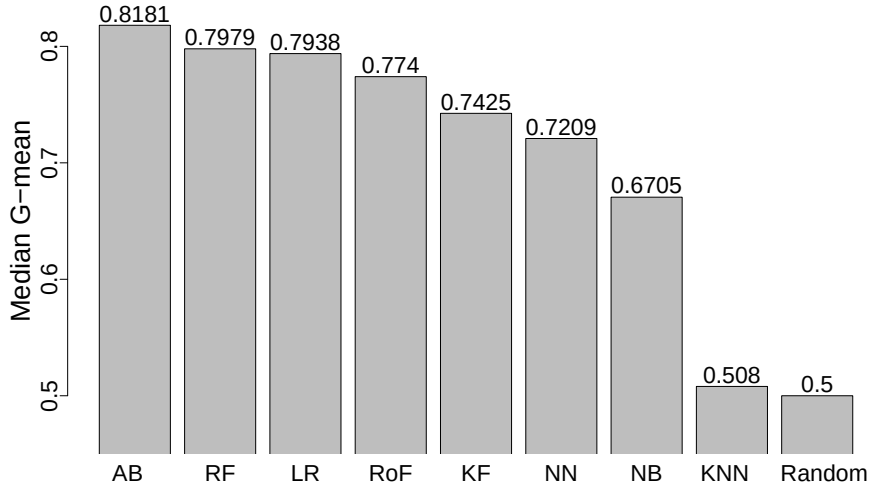


Figure 3.4: $5 \times 2cv$ *median G-mean*

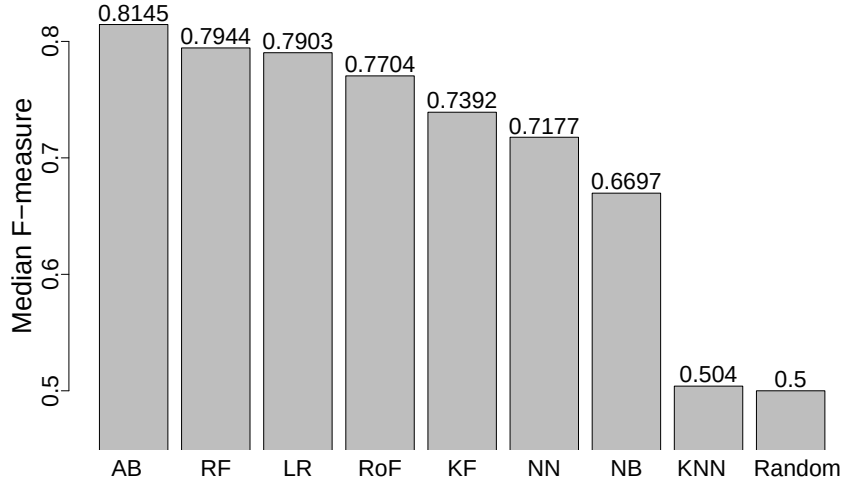


Figure 3.5: $5 \times 2cv$ *median F-measure*

Table 3.5: Average ranks of the folds (smaller is better)

	LR	RF	AB	KF	NN	RoF	KNN	NB	Friedman statistic (7)
Accuracy	**2.60**	**2.55**	**1.00**	5.20	5.60	4.05	8.00	7.00	66.93, $p < 0.001$
AUC	**2.50**	**2.50**	**1.10**	5.10	6.90	**3.90**	8.00	6.00	67.57, $p < 0.001$
G-mean	**2.55**	**2.55**	**1.00**	5.20	5.60	4.10	8.00	7.00	67.10, $p < 0.001$
F-measure	**2.55**	**2.55**	**1.00**	5.20	5.60	4.10	8.00	7.00	67.10, $p < 0.001$

Table 3.6: Cross-validated median IQR

	LR	RF	AB	KF	NN	RoF	KNN	NB
Accuracy	0.0037	0.0043	0.0016	0.0039	0.0052	0.0035	0.5812	0.0063
AUC	0.0021	0.0038	0.0020	0.0040	0.0176	0.0040	0.0534	0.0062
G-mean	0.0169	0.0196	0.0074	0.0176	0.0236	0.0155	0.0346	0.0173
F-measure	0.0168	0.0195	0.0074	0.0175	0.0235	0.0157	0.0374	0.0150

we identify two types of classifiers. The first type includes classifiers that have a critical difference with adaboost smaller than 2.9469 and are hence not significantly different (in bold). The second type are algorithms that perform significantly worse than the top performer. For the accuracy, G-mean and F-measure we note that RF and LR are not significantly different from AB and KF, NN, RoF, KNN and NB have significantly worse performance than AB. However, for the AUC we note that RoF likewise is not significantly worse than AB. The reason for this discrepancy could be found in the fact that the AUC considers the whole range of cutoff values whereas accuracy, G-mean and F-measure only consider a cutoff corresponding to a proportion of 10%. This implies that RoF performs well when considering the entire range of cutoff values, but performs slightly worse for the top 10%.

As a measure of dispersion, we report the interquartile range (IQR) of our algorithms across the folds (Table 3.6). The IQR ranges from 0.16% to 58% for accuracy, from 0.21% to 5.3% for AUC, from 0.75% to 3.46% for G-mean and from 0.74% to 3.74% for F-measure. This means that all algorithms produce stable results, except for KNN in the case of accuracy. Overall, k-nearest neighbors and neural networks produce the least stable results. Furthermore, the results in Table 3.6 reinforce the dominance of AB as the top performing algorithm since it has the lowest IQR across all performance measures. Also the IQR tends to favor LR over RF.

3.4.2 Disaggregated features

3.4.2.1 Information-fusion sensitivity analysis

In order to determine the top drivers in romantic tie prediction, we calculated the variable importances (or sensitivity scores) by means of information-fusion given by Eq. 3.12. Our α values are are the median 5x2cv AUCs (see Figure 3.3). The advantage of this technique over the traditional variable importance measure is that we use information of all prediction models [196]. Moreover, prediction models with higher AUC receive more weight in the fusion operator. In order to calculate the variable importance, we computed the median 5x2cv decrease in AUC for each model. The general idea is that the more a variable is associated with the response, the more predictive power it has, the more sensitive the model to a change in this variable is and the higher its mean decrease in AUC will be [123]. The final sensitivity score is calculated by inserting the α values and the median 5x2cv variable importances into Eq. 3.12.

Figure 3.6 shows the sensitivity score of all the predictors in decreasing order (solid line). We also added the cumulative percent of the raw sensitivity scores (the dashed line in Figure 3.6) in order to do a pseudo-Pareto analysis [175]. We notice a break in the solid curve at rank 18. Hence, predictors with a higher rank only have a marginal influence on the predictive performance of our model. Furthermore, our cumulative percent curve informs us that with 18 predictors (37 % of all the predictors) almost 90% of the cumulative sensitivity can be explained. We also notice that the traditional 80/20 Pareto rule holds in our case: with 10 predictors (i.e., 20% of all variables) we can explain 80% of the cumulative sensitivity score. Table 3.7 provides a list of the top 18 features. REC stands for recency, COUNT for frequency, and IND for indicator. We also added a column specifying the type of feature namely socio-demographic similarity (*D*), personal preference similarity (*P*), disaggregated frequency (*F*) and disaggregated time-related (*T*) communication features.

The results indicate that 44% of the top predictors are frequency variables, followed by time-related variables (28%), socio-demographic similarity (17%) and personal preference similarity variables (11%). This supports our hypothesis that time-related variables are important predictors of romantic ties. Looking at the communication type, comments are clearly the most important for both frequency and time-related variables, followed by tags and likes. This is in line with the results of Burke and Kraut [40] who state that more personal communication such as commenting has a stronger influence on social ties than passive communication such as liking. Liking on comments can also be seen as a more personal form of liking. A friend must first read through the comments on a user's posts before liking it, hence liking a comment is more time-consuming and thus more personal. Tagging is also a form of personal communication, because it informs us whether

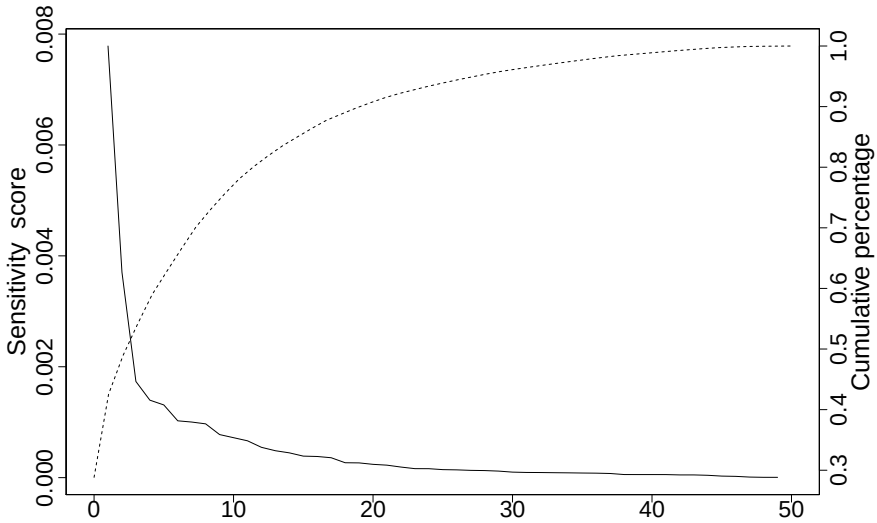


Figure 3.6: Scree plot of predictors

ego and alter communicate in real-life. These findings are in line with social signaling theory, which posits that the more time people invest in interacting with an alter, the higher the probability will be that a strong tie manifests itself [202]. The most important variables are related to socio-demographic similarity such as the gender and age difference.

Focusing on post types, we find that photos, statuses, videos and albums are key for both frequency and time-related features. This confirms the theory that videos, photos and albums are more dynamic and lively and thus induce more interaction [61]. Following social signaling theory, we can also argue that albums and videos demand more effort of the user to consume and therefore they serve as good indicators of strong ties. Other important post types are statuses.

To summarize, we extend the current theories on social tie prediction with disaggregated frequency and time-related variables. We found that variables that score high in terms of effort (e.g., commenting or tagging) and vividness (e.g., photos and videos) are especially important. Following the dimensions of Granovetter [102], we can state that disaggregated frequency and time-related variables related to the intimacy and intensity of the relationship are the most important dimensions in romantic tie prediction.

Table 3.7: Information-fusion based sensitivity score

Rank	Variable name	Median average sensitivity index	Type
1	IND (common gender)	0.0078	D
2	Age difference	0.0037	D
3	COUNT (tags on photos)	0.0017	F
4	IND (common location)	0.0014	D
5	COUNT (likes on statuses)	0.0013	F
6	REC (comments on statuses)	0.0010	T
7	REC (tags on photos)	0.0010	T
8	COUNT (common likes)	0.0010	P
9	COUNT (likes on photos)	0.0008	F
10	COUNT (common groups)	0.0007	P
11	REC (comments on photos)	0.0007	T
12	REC (tags on videos)	0.0005	T
13	REC (comments on albums)	0.0005	F
14	COUNT (tags on check-ins)	0.0004	F
15	COUNT (comments on photos)	0.0004	F
16	REC (comments on videos)	0.0004	T
17	COUNT (likes on photos)	0.0004	F
18	COUNT (comments on statuses)	0.0003	F

3.4.2.2 Partial dependence plots

To uncover the true relationship between our predictors and the response, we use partial dependence plots (PDP). PDP visualize the relationship between a predictor and a binary response, while keeping all other predictors constant [95]. PDP are created as follows. First a fusion model is built based on Eq. 3.11. Second, for each distinct value v of a predictor x a new data set is created. In this novel data set, x can only take this value v while all other variables are left untouched. Third, for each new data set the fusion model is deployed to score our response. Fourth, we take the mean of half the logit of the scores to obtain a single value p for all observations. Finally, the different values v of x are plotted against their respective value p [28]. In Figure 3.7 we provide the 5x2cv PDP for the top 10 predictors[2].

The top predictor in our model is whether or not alter and ego have the same gender (Figure 3.7a). We note a negative relationship between having the same gender and the probability to be in a relationship. This is logical since most of the relationships today are between people of the opposite sex. The second most important predictor is the age difference between ego and alter (Figure 3.7b). We see that there is an increase in probability with an age difference of approximately

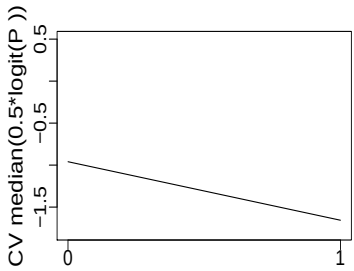
[2] All time-related variables are expressed as number of days

3 years (both in the positive and the negative sense). For the total number of photo tags (Figure 3.7c) we observe a positive influence on being a significant other. The same relationship can be observed with the number of status likes (Figure 3.7e) and the number of photo likes (Figure 3.7i). This is confirmed by the Facebook data science institute who noticed a rise in communication prior to installing the relationship [171].

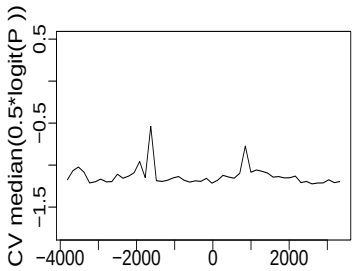
Next, we see that having a common location has a positive effect on being a significant other. Common location refers to whether or not ego and alter have indicated the same location as their hometown. Since couples mostly live together, egos mostly have the same location as their significant other. Next we note a positive relationship between the recency of comments on statuses (Figure 3.7f) and the recency of photo tags (Figure 3.7g) and the propensity of being a significant other. In other words, the longer it has been, the higher the propensity of being in a romantic relationship. A study of the Facebook data science institute revealed that after users changed their relationship status their interaction on Facebook gradually decreases [171]. The reasoning is that couples spend more time together and replace online interaction with more offline interaction. This theory is confirmed by the study of Burke and Kraut [40] who found that Facebook interaction is more common for people who do not communicate face-to-face. The couples in their study explicitly cited that they see each other every day and don't use Facebook to keep in contact. This explains why we observe a positive relationship between the recency variables and the probability of being a significant other. Finally, we observe a positive influence of the number of common likes and common groups and being a significant other. This can be explained by the theory of homophily:'birds of a feather flock together' [160]. Hence, people who are more alike and share common interests, have a higher chance of being in a relationship [208].

3.5 Conclusion

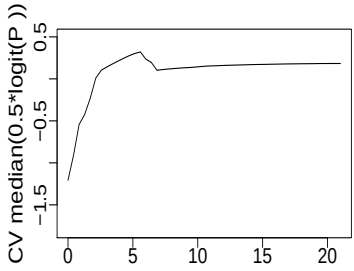
In this study we assessed the value of disaggregated time-related features in romantic tie prediction. Time-related variables are operationalized as the time since last communication or recency. Disaggregated features are defined as separate predictors measuring comments, likes and tags an ego has placed on an alter's albums, check-ins, locations, photos and status updates. Furthermore, we also added disaggregated frequency, socio-demographic and preference variables. We used four single classifiers (i.e., k-nearest neighbors, naive bayes, logistic regression and neural networks) and four ensemble classifiers (i.e., random forest, adaboost, rotation forest and kernel factory) to estimate being a significant other. Furthermore, we controlled for class imbalance by applying random oversampling, random undersampling and synthetic minority oversampling. We also applied information-fusion based sensitivity analysis to gain insight in the top drivers of romantic ties



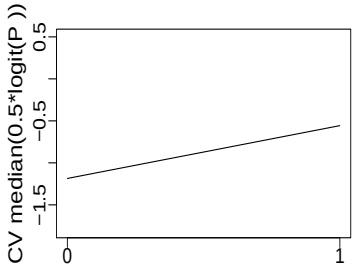
(a) Common gender



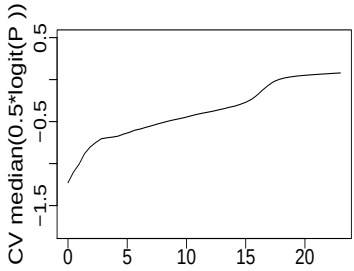
(b) Age difference



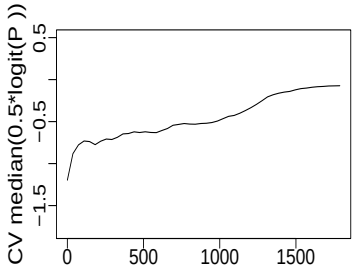
(c) Number of photo tags



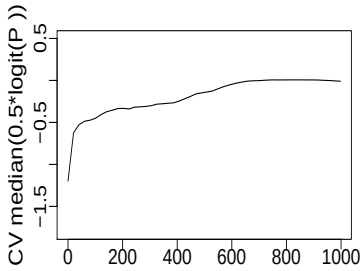
(d) Common location



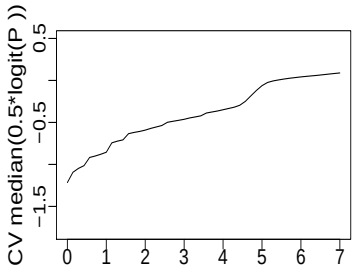
(e) Number of status likes



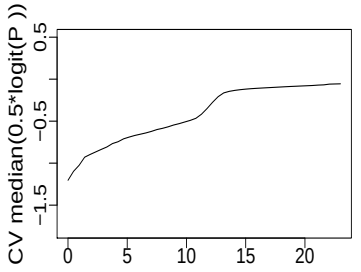
(f) Recency of comments on statuses



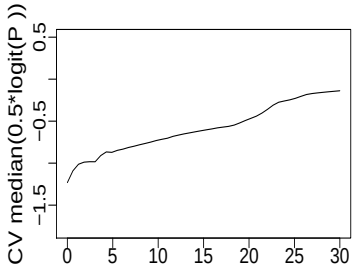
(g) Recency of tags on photos



(h) Common likes



(i) Number of photo likes



(j) Common groups

Figure 3.7: Partial dependence plots

and built partial dependence plots of the fusion model to study the form of the relationships that govern the model.

Our results indicate that we can predict romantic ties with high predictive performance. We found a median 5x2cv accuracy of up to 95.89%, an AUC of up to 97.56%, a G-mean of up to 81.81% and a F-measure of up to 81.45%. These results are in line with earlier studies that predict social ties [125, 126]. In terms of accuracy, AUC, G-mean and F-measure adaboost was the top performing algorithm, followed by random forest, logistic regression, rotation forest, kernel factory, naive Bayes, neural networks and k-nearest neighbors. We note that the top three algorithms performed equally well in statistical terms. All other algorithms performed significantly worse than adaboost. In terms of AUC, and in addition to random forest and logistic regression, rotation forest was also not significantly different from adaboost. This difference can be due to the fact the AUC is a portmanteau measure and thus considers all possible threshold values whereas the others have a threshold equivalent to a proportion of 10%. Hence, when speed is of an issue, one should consider random forest since it allows parallelization [18, 192]. The majority of the top predictors of romantic ties are socio-demographic, frequency and time-related variables related to commenting, liking and tagging. We found that socio-demographic similarity (e.g., having the same location or the same age category) and preference similarity have a positive effect on the chances of being in a relationship. This finding provides evidence of the presence of homophily when people establish a relationship [160]. In terms of interaction types we conclude that comments, likes and tagging are the most important, whereas photos, albums and videos are the most important post objects. Frequency and time-related variables that combine a high amount of time effort (e.g., commenting and tagging) with a high level of entertainment (e.g., albums and videos) are especially important, since they represent an intense and intimate relationship [102]. We found a positive relationship between the time since last interaction and being a significant other. Conversely, for the disaggregated frequency variables we observed a positive relationship with being a significant other. The explanation goes as follows. Prior to installing their relationship on Facebook couples interact more on Facebook. After establishing the relationship, couples then change Facebook communication into face-to-face interaction [40, 171]. This finding reinforces the fact that Facebook interaction is mostly used for keeping up with infrequent contacts [144].

Overall our results extend current theories on social tie prediction. In that regard, we provide important insight in the relationship of disaggregated time-related variables and social ties (i.e., significant other). In general, we show that an incorporation of disaggregated variables is thus necessary to discover the true effect on romantic partnerships, which is otherwise averaged out. Therefore, we suggest that disaggregated time-related variables be included in models since they

offer a deeper understanding in the relationship between ego and alter.

3.6 Practical implications

Our results provide important insights for researchers and academics involved in modeling social ties. Our research shows that including disaggregated time-related variables provides important insight in the true relationship between ego and alter and its predictors. Hence academics that want to conduct studies related to social tie prediction should include these variables. For example, researchers that are studying the existence of social circles in online social networks [e.g., 75, 229] could incorporate our time-related variables in their models and verify whether their findings still hold. Another good application of our time-related variables can be found in studies related to social diffusion [see 5]. These papers include proxies for tie strength to investigate the difference between strong and weak ties. Our results point out that time-related variables are important top predictors of social ties and hence have to be included to discriminate between strong and weak ties. In Table 3.7 we included a list of the top predictors of romantic ties, which can serve as a first indication.

Furthermore we also provided important insight on the form of the relationship between disaggregated time-related variables and romantic partnership. Our findings show that the general idea that a low recency leads to a higher chance of being in a relationship does not hold for romantic partners. On the contrary, we observe a positive relationship between the time since last communication and the probability of being a significant other. We included diagnostic partial dependence plots and theoretical justification to explain the observed relationship.

3.7 Limitations and future research

First, our study is limited since we do not include features related to the whole social network (i.e., topological features) in our analysis. Nevertheless, these variables have proven to be successful in predicting romantic relationships [14, 208]. Unfortunately we were unable to collect these features as they are very hard to obtain. In addition we chose not to include these features since we wanted to put emphasis on the characteristics that drive the interaction between ego and alter and how this influences their relationship. Instead of focusing on properties of the network, we thus decided to focus on the properties of the friendship between ego and alter [114].

Second, we were not able to include private communication or messages since Facebook and privacy regulations do not allow us to extract this information. We focused on observable interaction between ego and alter: our disaggregated fre-

quency and time-related variables capture the whole spectrum of the observable interaction on Facebook.

Third, our dependent measure is whether or not an ego is an alter's significant other. Hence, we do not study the traditional tie strength problem. Other studies that have studied tie strength in social media [e.g., 100] directly ask the user to rate the strength of their relationship. A possible avenue for future research could be to investigate whether our time-related variables remain the top predictors of tie strength and whether the relationship changes when using other definitions for tie strength.

Fourth, since our dependent measure is whether or not an ego is an alter's significant other we are in fact modeling social ties. However, a lot of different social ties exist on Facebook next to a significant other (e.g., family relationships or colleagues). An interesting avenue for future research could be to control for the different types of social ties and see whether the results hold for different kinds of ties.

As a final note we want to say that although our data have shortcomings, this is the first study to provide such an elaborate analysis using Facebook data. We think this is because these data are very hard to obtain. Therefore we are confident that this study makes a significant contribution to literature.

3.8 Appendix

Median 5x2cv Accuracy, G-mean, F-measure and AUC per algorithm and per data sampling technique.

Sampling technique	Performance measure	LR	RF	AB	KF	NN	RoF	KNN	NB
RUS	Accuracy	0.9491	0.9542	0.9548	0.9420	0.9455	0.9469	0.8560	0.9211
	G-mean	0.7736	0.7944	0.7992	0.7412	0.7574	0.7618	0.5108	0.6596
	F-measure	0.7702	0.7909	0.7957	0.7379	0.7540	0.7583	0.5002	0.6578
	AUC	0.9669	0.9663	0.9741	0.9430	0.9596	0.9644	0.8174	0.9309
SMOTE	Accuracy	0.9485	0.9506	0.9539	0.6131	0.9458	0.9443	0.8990	0.9198
	G-mean	0.7709	0.7788	0.7952	0.4763	0.7587	0.7520	0.5680	0.6647
	F-measure	0.7675	0.7752	0.7917	0.3774	0.7554	0.7487	0.5648	0.6646
	AUC	0.9682	0.9607	0.9747	0.9389	0.9546	0.9419	0.8439	0.9227
ROS	Accurcay	0.9536	0.9545	0.9589	0.9423	0.9375	0.9503	0.8664	0.9237
	G-mean	0.7938	0.7979	0.8181	0.7425	0.7209	0.7740	0.5080	0.6705
	F-measure	0.7903	0.7944	0.8145	0.7392	0.7177	0.7704	0.5040	0.6697
	AUC	0.9686	0.9682	0.9756	0.9443	0.8961	0.9616	0.7605	0.9336

4

Comparing the Ability of Twitter and Facebook Data to Predict Box Office Sales [1]

Abstract

This paper aims to determine which social media platform (Facebook or Twitter) is most predictive of box office sales, which algorithm performs best, and which variables are important. To do so, we introduce a holistic social media analytical approach that consists of two stages. In the first stage, we compare several models based on Facebook and Twitter data. We benchmark these model comparisons over various data mining algorithms (i.e., linear regression, k-nearest neighbors, decision trees, bagged trees, random forest, gradient boosting, and neural networks) using five times two-fold cross-validation. In the second stage, we apply information-fusion sensitivity analysis to evaluate which variables from which platform and from which data type are driving the predictive performance. The analysis shows that Facebook data outperform Twitter data at predicting box office sales by more than 11% in RMSE, 13% in MAE, 14% in MAPE, and 47% in R^2. In terms of the best algorithm, we find that random forest is the top performer

[1] Based on: *Bogaert, M., Ballings, M., Van den Poel, D., & Oztekin, A. (2018). Comparing the Ability of Twitter and Facebook Data to Predict Box Office Sales. Under revision in the Journal of Management Information Systems.*

across all performance measures. We also find that including user-generated content (UGC) does not significantly increase the predictive power for both Twitter and Facebook data. Our sensitivity analysis reveals that the number of Facebook page likes (i.e., a page-popularity indicator) is the most important variable, followed by the hype factor of Facebook comments (i.e., user-generated content). The results also show that volume is a better predictor than valence. Our results provide clear guidelines for practitioners, marketers and academics who want to model box office sales using social media data.

4.1 Introduction

Since the rise of social media, substantial research has been conducted on the relationship between social media and movie sales [68]. Data originating from social media sites such as Yahoo! Movies and Twitter have been used to explain this relationship [72, 193]. 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 [46]. Asur and Huberman [10] 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 [193].

While research concerning social media and box office sales has advanced to some extent, it still suffers from two 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 sales [153]. Therefore, it is important to know how both data sources compare in predicting box office sales. Since both Facebook and Twitter generate considerable amounts of data, collecting and parsing data from both data sources can be intractable. 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 (i.e., Facebook and Twitter pages) impact consumer purchase behavior [101]. Besides UGC and MGC, Facebook and Twitter pages also contain page-popularity indicators (PPI) (e.g., the number of Facebook page likes or Twitter page followers) that are indicative of movie sales [173]. To the best of our knowledge, no study has collectively included UGC, MGC and PPI in their box office predictions and conducted a comprehensive analysis of Facebook and Twitter. Hence, this leaves two important questions

unanswered: (1)'Which social media platform is the most predictive of box office sales?', (2) 'Which algorithm performs the best?', and (3) 'Which variables are the most important drivers of box office sales?'.

This paper contributes to literature in several ways. First, we predict box office sales for 231 movies using data from both Facebook and Twitter. To conduct a fair comparison between both platforms, we build two models for each platform. One model is based on MGC and PPI such as the total number of posts or tweets and the total number of page likes and followers. The second model augments the first model with UGC such as total number of comments and the total number of replies. To ensure that our results are robust, we benchmark these model comparisons over 7 state-of-the-art algorithms: regularized linear regression, k-nearest neighbors, decision trees, bagged trees, random forest, gradient boosting and neural networks. Finally, we apply information-fusion sensitivity analysis to evaluate which variables from which platform (i.e., Facebook or Twitter) from which data type (i.e., UGC, MGC and PPI) are the driving force of predictive performance. To demonstrate our contribution, we introduce a social media analytical methodology, which is an enhancement of the CRISP-DM framework [43]. 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. 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, MGC and PPI variables.

The remainder of this paper is organized as follows. First, we provide an overview of the existing literature. Second, we discuss our methodological framework, the extracted data, the variables and the algorithms. Third, we describe our results and practical implications. To conclude, we elaborate on the limitations and avenues for future research.

4.2 Literature overview

Research on box office predictions has mostly studied two types of variables: movie characteristics and UGC, such as WOM [130]. The former includes variables such as the cast of the movie, the content of the movie and the release time of the movie [141]. The latter consists in the influence of WOM volume and valence on movie sales [72]. The majority of the literature on box office predictions using WOM has focused on Yahoo!Movies [46, 62, 167], blog posts [164], Google searches [152], and IMDB [151]. Most of these studies conclude that online chatter has a significant influence on movie sales over and above other data sources, such as ratings [72]. 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 Face-

book. The reasons for this shift are manifold. First, Facebook and Twitter have a large user base with respectively 2.01 billion [84] and 328 million monthly active users [210]. 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) [83]. Third, Facebook and Twitter pages are updated frequently and have a standardized API which allow for a user-friendly and structured data collection [161]. Finally, both platforms contain a lot of user-created and company-created content as well as page popularity indicators that have proven to have a significant impact on movie sales [68]. For the aforementioned reason, we decide to focus our study on Facebook and Twitter [2].

Studies on social media data and box office revenues can be categorized according to several dimensions: (1) whether they use Facebook or Twitter, (2) whether they include PPI, MGC and UGC, (3) whether they compare the predictive performance of both platforms and (5) the number of movies they predict (Table 4.1). Studies including Twitter use the tweets about a movie to forecast movie sales [10]. For example, Rui et al. [193] 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 [173]. For example, Oh et al. [173] discovered a positive relationship between the number of likes for a certain movie and gross box office revenues. Page popularity indicators (PPI) refer to meta-information of a Facebook or a Twitter page such as the page likes and the number of followers [173]. For example, Ding et al. [68] 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 [101]. 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 [78]. For example, Asur and Huberman [10] use both the rate of the tweets (i.e., volume) and two sentiment measures (i.e., polarity and subjectivity) to model box office revenues. Finally, studies comparing social media platforms assess which data hold the most predictive power. For example, Oh et al. [173] found, using a robust OLS model, that Twitter lost all predictive power of movie sales when Facebook data were entered in the model. However, they

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

include the volume of WOM (e.g., the total talk on Facebook and the total tweets on Twitter) and PPI (e.g., total page likes and the number of followers) but neglect to add MGC as well as the valence of the online chatter.

Table 4.1 summarizes the literature on movie sales and social media data. From Table 4.1, it is clear that no study has included PPI, MGC and UGC to predict box office sales and conducted an analysis of Facebook and Twitter (i.e., the dominant platforms in the marketplace). Furthermore, we include more movies than any other study to date. The question whether or not Facebook or Twitter is the most important platform in predicting box office sales is non-trivial. Facebook and Twitter are becoming more and more important as a tool for building brand equity and increase consumer engagement [60, 155]. Applications such as Facebook Ads even allow firms to target specific audiences with their brand posts [83]. 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 [85, 211]. 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 which platform to choose. Once a firm has chosen its preferred platform, a second question is which variables and what type of data to gather and which data has the highest impact on predictive performance. UGC and PPI have proven to have a significant impact on box office sales predictions [68, 130]. 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 [60, 101, 134]. Hence, amongst the clutter of UGC, MGC and PPI marketers want to know on which type of data to focus. 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 4.1: Overview of box office prediction literature including Twitter and/or Facebook

Study	Platform	PPI	MGC		UGC		Compare platforms	N(Movies)
			Volume	Valence	Volume	Valence		
Asur and Huberman [10]	Twitter				X	X		24
Reddy et al. [187]	Twitter				X	X		1
Wong et al. [220]	Twitter					X		34
Apala et al. [3]	Twitter					X		35
El Assady et al. [78]	Twitter					X		20
Guàrdia-Sebaoun et al. [104]	Twitter				X	X		32
Jain [121]	Twitter				X	X		30
Rui et al. [193]	Twitter				X	X		63
Arias et al. [6]	Twitter					X		50
Du et al. [71]	Twitter [2]				X	X		24
Hennig-Thurau et al. [111]	Twitter				X	X		105
Liu et al. [149]	Twitter [3]				X	X		57
Gaikar et al. [97]	Twitter				X	X		14
Kim et al. [130]	Twitter [4]				X	X		212
Ding et al. [68]	Facebook	X						64
Oh et al. [173]	Twitter, Facebook	X			X		X	106
Baek et al. [15]	Twitter [5]				X			145
Our study	Twitter, Facebook	**X**	**X**	**X**	**X**	**X**	**X**	**231**

[2] The data in their study are collected from the Chinese microblogging site Weibo, a Chinese counterpart of Twitter.

[3] The data in their study are collected from the Chinese microblogging site Weibo, a Chinese counterpart of Twitter. They also perform a within platform comparison between their approach and the method of Asur and Huberman [10]. Since this is within the same platform, we not see this as comparing different social media platforms.

[4] The data in their study were obtained from pulseK, which 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.

[5] 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. Next to UGC and PPI, we also include MGC, such as the number (and valence) of posts and the number (and valence) of tweets generated by the firm itself. 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 two types of models for each platform. The first type only uses PPI and MGC. The second type adds UGC to the first type. The reason is that UGC induces a large computational overhead, and should only be collected when it significantly improves predictive performance. Thus, in addition to analyzing the predictive ability of Facebook and Twitter, we also assess the added value of UGC. A previous study by Oh et al. [173] concluded that Twitter follows had a significant positive influence on movie sales when studied in isolation. However, when Facebook likes were introduced to the model, Twitter follows 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 [153] 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 and MySpace. 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 over 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). Moreover, we also determine which algorithm has the best performance. Research has shown that machine-learning algorithms improve the performance of box office predictions [198]. However, one major downside of these social media studies is that they neglect to use ensembles methods [6, 149]. Nevertheless ensemble techniques have proven to be superior in movie success prediction [141] and in other social media applications, such as event prediction [32] and daily sales forecast [53].

The second stage summarizes the information from both platforms and prediction models with a technique called information-fusion sensitivity analysis [176]. 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 [55]. 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 certain variables related to page popularity and WOM (or more in general UGC) would be of major importance in comparison to MGC. Page popularity indicators (e.g., the number of page likes on Facebook and the number of Twitter followers) show

a degree of interest towards a certain movie. According to the personal consumer engagement behavior theory, liking or following a movie reflects intrinsic motivation and involvement, and hence leads to higher movie sales [173]. Other authors argue, based on social impact theory, that more Facebook likes represent a higher social size and conversely higher social impact [68]. Therefore, more Facebook likes lead 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 [151]. The former states that people can only consider products of which 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 [193]. The latter helps people create their attitude and opinions towards the product through the information they receive, which in turn affects their purchase decision [72]. As the valence (or sentiment) of the tweet becomes significantly more positive, the persuasive effect becomes larger. 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 [151] 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. [220] came to the same conclusions. On the other hand, Chintagunta et al. [46] found that valence, and not volume, was the most important variable. Rui et al. [193] found that both volume and valence had an effect on box office revenues. Hennig-Thurau et al. [111] 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 [10] used a simple positive and negative tweet ratio for valence, while Kim et al. [130] employs the total number of emotional, positive and negative SNS mentions.

Current research concerning social media and box office sales did not investigate 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 [60], customer spending, cross-buying and profitability [134]. Compared to UGC, Goh et al. [101] 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 conclusion, we are the first to conduct such an extensive analysis of Face-

book and Twitter within the context of box office sales. We present theoretical evidence why we believe Facebook would be more indicative of movie sales. Moreover, we combine all PPI, UGC and MGC (both volume and valence) that have been proposed in literature and assess their predictive power. Since previous literature neglects to include all variable types in one study, we are the first to gain insight in the relative importance of PPI, UGC and MGC in the context of box office sales. On the theoretical side, we contribute to literature by assessing which content type dominates box office revenues while using social media data. On the methodological side, we contribute to literature by analyzing the largest collection of movies with the widest range of algorithms to date. Moreover, we introduce a generic social media analytical framework that can help researchers and practitioners replicate our methodological approach in other similar settings. In the next section, we discuss our framework, materials and methods.

4.3 Methodology

4.3.1 Framework

The framework employed in our study is a holistic integration of the well-known CRISP-DM methodology [43]. The CRISP-DM framework is the most commonly used methodology in analytics and ensures robust results [176]. 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 sensititvity analysis. In our framework the three data sources (i.e., BoxOfficeMojo, Facebook and Twitter) are crawled from the internet using the API. Figure 4.1 summarizes the proposed social media analytical framework.

The first step is the data collection. In this step the data is gathered from the Twitter and Facebook API and the BoxOfficeMojo website for the desired movies. The second step involves the inspection of the raw data sources (i.e., data understanding). The third step involves cleaning, handling and merging of the Twitter and Facebook data sources. There are two different data preparation procedures. The first one involves numeric and time variables that do not require any text processing. The second one includes text and sentiment analysis. The output of this step is several basetables including user-generated and page-generated content of both platforms (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 Facebook and Twitter variables. Using the fusion model, variable importances are assessed to uncover the driving forces of

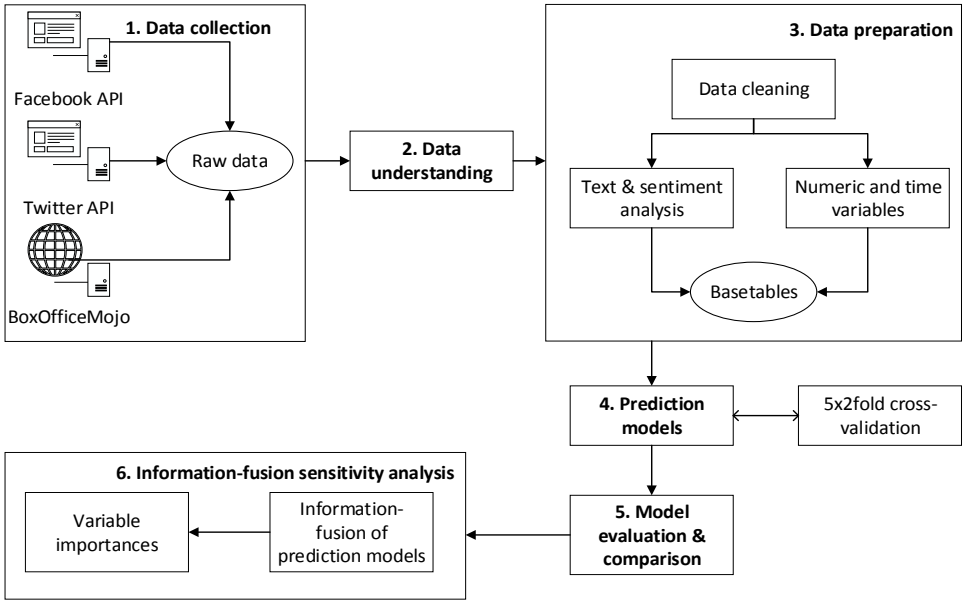


Figure 4.1: Social media analytical framework

predictive performance.

4.3.2 Data

We extracted data from 231 movies released between January 2012 and December 2015 that had a verified[3] Facebook or Twitter page. We obtained data from the Facebook and Twitter pages from the start of their very existence until the time of collection (August 2016). We note that 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 [85, 211]. The extracted data is available in JSON format, which allows for easy and fast processing [161]. Movie sales data (i.e., gross box office revenues) were collected via BoxOfficeMojo within the same time window [33].

The same data types were collected for both platforms and can be categorized into PPI, UGC, MGC. Page popularity indicators (PPI) are overall performance indicators of a Facebook or a Twitter page. For example, the total number of page likes and total amount of social media buzz (i.e., talking about) are the most important PPI on Facebook. MGC refers to all information on a Facebook page or Twitter wall created by the page owners (in our case movie producers). 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 content (UGC) consists of the Facebook comments and Twitter replies[4] posted by other users and the interactions on those comments and replies [197]. For example, the average sentiment of replies or comments and the number of likes on a post, comment or reply are user-generated content. Table 4.2 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 231 movies. For example, we collected 95,725 posts and 150,599 tweets in total across all movies and the median number of marketer-generated posts across all movies is 246 with a minimum of 23 posts and a maximum of 2280 posts for a movie. The final row of Table 4.2 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 [79, 177, 193].

Based on these three categories, we propose 4 data sets to compare both platforms. The baseline models include PPI and MGC of Facebook and Twitter. The augmented models include PPI, MGC and UGC. This is motivated by the fact that collecting UGC induces a large computational overhead. For example, on Face-

[3]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.

[4]We filtered out replies that did not respond to a specific Tweet on the timeline.

Table 4.2: Descriptive statistics

Type of content		Total	Median	Min	Max
MGC					
	Facebook	95,725	246	23	2280
	Twitter	150,599	432	8	2992
UGC					
	Facebook	2,966,421	6120	127	243,008
	Twitter	1,596,501	4079	130	69,920
PPI					
	Facebook	-	247,000	1142	11,280,000
	Twitter	-	5834	127	616,600
Revenues ($)		-	24,480,000	25,480	356,500,000

book the API allows to easily collect the wall of a movie page. If you want to collect the comments (UGC), to sequence over all the collected wall posts and scrape them individually. 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 4 models we also added two models that combine both Facebook and Twitter data, one with UGC and one without UGC. These models aim to check whether the UGC has added value across both platforms and to evaluate the best algorithm. Table 4.3 summarizes the models used in this study including the data sources from which they are composed of and the total number of variables. We note that for every Twitter variable, we tried to make a Facebook counterpart. In some cases it was not possible to recreate the variables, this explains the small difference in the number of variables.

4.3.3 Variables

Table 4.4 provides an overview of our predictors, together with an example for each platform. 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, even though other researchers have included other variables related to the movie itself such as number of screens and the genre [68]. We deliberately exclude these variables from our analysis since we are only interested in the predictive power of social media data and how they compare against each other.

The variables in Table 4.4 are categorized according to the data sources identified in Section 4.3.2. PPI only encompass volume measures, whereas MGC and UGC are WOM measures and contain volume and valence or a combination of

Table 4.3: Overview models

Models	Facebook			Twitter			N(variables)
	PPI	MGC	UGC	PPI	MGC	UGC	
Fb:base	X	X					55
Fb:plus	X	X	X				104
Tw:base				X	X		58
Tw:plus				X	X	X	106
FbTw:base	X	X		X	X		113
FbTw:plus	X	X	X	X	X	X	210

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

both. For example, the number of positive posts is classified as a combination of volume and valence, whereas the percentage of positive posts and the average sentiment of the post is only valence. Both volume and valence are further divided into unrestricted and time-restricted variables. The former have no time component and their value is fixed at the time of scraping (i.e., aggregated over the whole time period until August 2016). The latter compare the creation date of the object (e.g., a post, tweet, comment or reply) in relation to the release date of the movie.

The first category are the more general PPI, such as the number of followers of the movie on Twitter and the number of movie page likes on Facebook [173]. For example, Ding et al. [68] concluded that a 1% increase in the number of page likes one week prior to release results in a 0.2% increase in box office revenues in the opening weekend. A comparable study on Twitter was done by Rui et al. [193]. They conclude that having more followers on Twitter influences box office revenues. Before going deeper into MGC and UGC, we elaborate the text analysis and sentiment analysis part of our approach which is used to calculate valence variables.

Table 4.4: Overview variables

Category			Facebook	Twitter
PPI			Number of Likes	Number of followers
MGC	Volume	Unrestricted	Number of posts	Number of tweets
			Mean number of post per day	Mean number of tweets per day
		Time-restricted	Number of posts before release	Number of tweets before release
			% of posts 1 week before release	% of tweets 1 week before release
	Valence	Unrestricted	% positive posts	% positive tweets
			Mean sentiment score posts	Mean sentiment score tweets
		Time-based	% neutral posts 2 weeks after release	% neutral tweets 2 weeks after release
			Change post sentiment before/after release	Change tweet sentiment before/after release
	Combination	Unrestricted	Number of positive posts	Number of positive tweets
			Number of negative posts	Number of negative tweets
		Time-restricted	Number of neutral posts after release	Number of neutral tweets after release
			Number of negative posts after release	Number of negative tweets after release
UGC	Volume	Unrestricted	Hype factor	Hype factor
			Mean likes on comments	Mean retweeted replies
		Time-based	Number of comments after release	Number of tweets after release
			% comments 1 week before release	% replies 1 week before release
	Valence	Unrestricted	Ratio positive/negative comments	Ratio positive/negative replies
			% positive comments	% negative comments
		Time-based	Sentiment comments before release	Sentiment replies before release
			% neutral comments 2 weeks after release	% neutral comments 2 weeks after release
	Combination	Unrestricted	Number of positive comments	Number of positive replies
			Number of neutral comments	Number of neutral replies
		Time-restricted	Number of neutral comments 2 weeks after release	Number of neutral replies 2 weeks after release
			Number of positive comments 1 week before release	Number of positive replies 1 week before release

4.3.3.1 Text and sentiment analysis

Before we could perform sentiment analysis, we had to perform some text cleaning: transforming the text to lower case, removing punctuation, leading and trailing white space, numbers and web links [118]. After the text cleaning, the sentences were broken down into words. To conduct sentiment analysis we perform the lexicon-based method, as this is considered the most common method to assess the influence of WOM in social media [127]. 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 [204]. We used an English lexicon with valence scores of 13,915 words, so only English sentences were considered. Next, the words in each sentence were matched the corresponding valence scores on a 9-point Likert scale. We rescaled valence-scores around 5 such that 0 corresponded to a neutral comments. Hence, the valences scores range from $[-4, 4]$, with a value of 0 reflecting a neutral, -4 a very sad and 4 a happy word. The final valence score is achieved by averaging across all the words in the text-item (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 [193]. 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[5]

4.3.3.2 MGC and UGC variables

In Table 4.4 MGC consists of all variables related to the posts and tweets posted by the page administrators themselves. This category can again be subdivided in five categories whether or not the variables are unrestricted or time-restricted, volume or valence or a combination of both. First, unrestricted volume variables describe the frequency of certain characteristics of a post or tweet over the entire period. For example, the number of tweets computes the sum of all tweets sent from the moment of creation of the page until August 2016. Another type of variable is the tweet-rate which represent the average number of tweets per day [10]. Time-related volume variables are frequency-based variables restricted to a certain time-window. We follow the recommendation of Ding et al. [68] and Kim et al. [130] and do not only include variables before and after release, but also one week prior and two weeks afterwards. Asur and Huberman [10] identify the period 1

[5]We tried several lower and upper-bounds to classify negative and positive words: -0.5 and 0.5, -0.5 and 1.5, and -1.5 and 1.5. For all three options we looked at their relative and absolute frequency of negative, neutral, and positive words. We picked the threshold -0.5 and 0.5 since this gave a sufficient amount of negative words and a balanced amount of neutral words.

week prior and two weeks after release 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 [10, 111]. For example, the number of posts before release aggregates the total number of posts before the release date, whereas the percentage of posts 1 week prior to release computes the ratio of the total number of posts one week prior to release and the total number of posts over the whole period. Unrestricted valence measures use sentiment analysis and are calculated over the entire time window. We included both the average sentiment score as well as a classification into positive, negative or neutral (see Section 4.3.3.1). For example, the average sentiment of a post or the percentage of positive tweets 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) [10]. Time-based valence measures are sentiment variables calculated in relation to the release date. For example, the percentage of neutral tweets two weeks after release computes the number of neutral tweets two weeks after release relative to the total number of tweets. Furthermore, we also included the change in sentiment scores before and after release [130]. This variable measures the effect of a movie being better or worse than anticipated. Finally, unrestricted and time-restricted combination measures calculate the frequency of sentiment classes over the entire period or a fixed period in time. For example, the total number of positive posts is an unrestricted combination and the total number of positive tweets before release is a time-based combination. We note that percentage of positive 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 on tweets and comments on posts on the official movie page. These variables are fairly similar to their MGC counterparts with a few additions. For example, as proposed by Gaikar et al. [97], we calculate the hype factor which measures how many distinct users have reacted to a certain post or tweet relative to the total number of comments or posts. If the hype factor comes close to 1, it implies that a lot of distinct users have reacted. If the hype factor is close to 0, it means that several users replied more than once.

4.3.4 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 [32] and have proven to yield good performance in movie sales predictions [141].

4.3.4.1 Regularized linear regression

We apply linear regression with lasso (i.e., least absolute shrinkage and selection operator) to control for overfitting [231]. The lasso restricts the absolute sum of the regression coefficients to a predefined value. As a consequence some of the regression coefficients are shrunken towards zero [109]. We used the R-package *glmnet* to build the prediction model [92]. We cross-validated the $nlambda$ parameter in terms of Root Mean Square Error (RMSE) by sequencing over all its values (default 100).

4.3.4.2 K-nearest neighbors

K-nearest neighbors (KN) is a non-parametric pattern recognition method which makes no assumptions about the form of the regression function [2]. We used the k-d tree algorithm to implement k-nearest neighbors [27]. In classification the value of a new instance is determined by taking the majority vote of the K most similar instances. In regression, the prediction of a new sample is the average value of the K nearest neighbors. Hence, it is important to cross-validate the value of K (i.e., the number of nearest neighbors to determine the final predictions). We iterated over all values from $K = \{1, 2, ..., 150\}$ to determine the optimal K in terms of RMSE. To implement KN we used the R-package FNN [29].

4.3.4.3 Decision trees

We use the Classification and Regression Tree (CART) approach by [37] using binary recursive partitioning to build a decision tree (DT). The algorithm is called binary since two child nodes are created from each parent node. The recursive nature stems from the fact that each child node will become a parent unless it is a terminal node. Regression trees divide the feature space in several distinct and non-overlapping regions. Predictions are made by taking the average value of the training instances of the region to which the new instance belongs [122, p. 306-311]. To avoid overfitting we pruned our regression tree by cross-validating the cost complexity parameter (cp). We sequenced over all values from $cp = \{0.001, 0.002, 0.003, ..., 0.199, 0.200\}$ and selected the value of cp that minimized the RMSE. We used the R-package $rpart$ to build our decision trees [205].

4.3.4.4 Neural networks

A neural networks (NN) is a non-parametric machine learning method which mimics the behavior of the human brain [212]. The algorithm consists of three hidden layers, namely the input layer, the hidden layer and the output layer. We use a feed-forward neural network optimized by BFGS with one hidden layer as our algorithm implementation [70]. To build our feed-forward neural network, we use

the R-package $nnet$ [189]. Several operations and parameter tuning are required to effectively implement a NN. First, all numerical predictors - binary variables are disregarded- are rescaled to $[0, 1]$ to avoid numerical and computational problems. Second, the starting weights are chosen at random [190, p. 154]. Third, we set the $entropy$, the $rang$, the $abstol$, the $reltol$, the $MaxNWts$ and the $maxit$ parameter to respectively the maximum conditional likelihood, 0.1, 1.0e-4, 1.0e-8, 1000 and 1000. Finally, we optimized the weight decay parameter ($decay$) and the number of nodes in the hidden unit ($size$) by performing a grid search across the values for $decay = \{5, 10, 20\}$ and $size = \{0.001, 0.01, 0.1\}$ [190, p. 163 - 170].

4.3.4.5 Bagged trees

Bagging tries to cope with high variance of decision trees by means of 'bootstrap aggregation' [34]. This implies that independent bootstrap samples of the same size as the training data are constructed by sampling with replacement. Consequently several decision trees are built using these bootstrap samples. Finally the different trees are aggregated by averaging the output. We used the R-package $ipred$ to build bagged CART trees with 25 bags ($nbagg$) [182].

4.3.4.6 Random forest

Random forest adds an additional layer of randomness to the bagging algorithm [35]. Randomized CART trees (i.e., the best split at each node is determined from a random subset of features) are grown on independent bootstrap samples. The multiple trees are aggregated by means of majority vote. To implement the algorithm only two variables have to be supplied: the number of random predictors to consider at each node split ($mtry$) and the number of trees ($ntree$). We set the $mtry$ parameter to the square root of the number of predictors (default) and the $ntree$ parameter to 500 [35]. We use the R-package $randomForest$ to implement the algorithm [147].

4.3.4.7 Stochastic gradient boosting

We use Friedman's gradient boosting machine to implement the boosting algorithm [93]. Gradient boosting sequentially adds weak learners in a greedy fashion such that the loss function is minimized at each iteration. CART trees are found to be a superior weak learner for boosting, since they are flexible, easily added together and very fast [133, p. 203-208]. We use the R-package gbm, which requires several tuning parameters [188]. First, the tree depth ($interactiondept$) and the number of observations in the terminals node ($nminobsinnode$) restrict the depth of trees to create a weak learner. Second, to make sure that the algorithm finds a global optimum and does not overfit, we employ shrinkage ($shrinkage$). Finally,

to avoid early stopping we set the number of iterations ($n.trees$) to a high number. We optimized the aforementioned parameters in terms of RMSE by performing a grid search across: $interactiondept = \{1, 3, 5, 7\}$, $shrinkage = \{0.01, 0.1\}$, and $ntrees = \{100, 500, 1000\}$ [133, p. 203-208].

4.3.5 Performance evaluation and cross-validation

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. Many researchers choose to include the adjusted R^2 instead of the normal R^2 [10, 68]. The main reason why we use the regular R^2 is that we do not calculate the R^2 on the training set, but on a separate hold-out test set. The main reason of adjusting the R^2 is to remedy the inflation of the training set performance caused by increasing the number of variables. Since we use a hold-out test set increasing the number of variables does not impact test set performance and our algorithms do not blindly maximize the R^2. As a result, there is no problem in using the regular R^2 [165]. The RMSE, MAE, MAPE, and R^2 are defined as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}_i|, \tag{4.1}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}, \tag{4.2}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i - \hat{Y}_i|}{Y_i}, \tag{4.3}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = 1 - \frac{SS_E}{SS_T}, \tag{4.4}$$

with N the number of observations (231), 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 [133, p. 95].

To make sure our results are robust we employ five times two-fold cross-validation (5x2cv) [65]. 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. If tuning of the hyper-parameters was necessary, the training set was split again in two equal parts. Afterwards, the full training was used to estimate the final model. 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 4.3), wins-ties-losses tables are constructed [64]. To test for significant wins-ties-losses we use the non-parametric Wilcoxon rank test [218]. We also adapt the p-values with Bonferroni-Dunn corrections to control for multiple comparisons and family-wise error [50]. To compare the performance of each algorithm against the top performer, we employ the non-parametric Friedman test with Bonferroni-Dunn post-hoc test [96].

4.3.6 Information-fusion sensitivity analysis

To uncover which variables are the driving force of 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 [174, 196]. 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:

$$\hat{y}_i = f_i(x_1, x_2, ..., x_n) = f_i(\mathbf{x}), \tag{4.5}$$

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:

$$\hat{y}_{fusion} = \Psi\left(\hat{y}_1, \hat{y}_2, ..., \hat{y}_7\right) = \Psi\left(f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_7(\mathbf{x})\right), \tag{4.6}$$

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. 4.6 becomes :

$$\hat{y}_{fusion} = \sum_{i=1}^{7} \omega_i f_i(\mathbf{x}) \quad where \sum_{i=1}^{7} \omega_i = 1. \tag{4.7}$$

The value of weighting factor ω_i is proportional to the relative predictive performance of prediction model $\hat{y}_i$. Hence, a lower forecasting error of model $\hat{y}_i$ will result in a larger weight ω_i in Eq. 4.7 and hence more influence in the calculation of the information-fusion model $\hat{y}_{fusion}$. In our case ω_i is determined using the RMSE:

$$\omega_i = 1 - \frac{RMSE_i}{\sum_{i=1}^{7} RMSE_i}. \tag{4.8}$$

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 [31]. The variable importance of a certain variable j is determined by permuting on that variable and re-deploying the prediction model using this permuted variable [215]. The difference in RMSE before and after permutation is variable importance of variable j. We repeat this process three times and the result is called the mean decrease in RMSE. In mathematical terms this becomes:

$$V_{i,j} = \frac{1}{3}\sum_{k=1}^{3} RMSE_{i,j}^{k} - RMSE_{i,base}^{k}, \tag{4.9}$$

with V_{ij} the importance of variable j in prediction model i, $RMSE_{i,j}$ is the RMSE of prediction model i with variable j permuted and $RMSE_{i,base}$ the baseline RMSE of prediction model i.

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 [55]. If we rephrase Eq. 4.7 in terms of importance of variable j with 7 prediction models, this becomes:

$$V_{fusion,j} = \sum_{i=1}^{7} \omega_i V_{i,j}. \tag{4.10}$$

The measure in Eq. 4.10 indicates how much the overall RMSE would increase if variable j would not be included in the model. We note that all of the aforementioned measures are 5x2cv cross-validated. Hence, Eq. 4.10 represents the median 5x2cv mean decrease in RMSE of variable j.

4.4 Results

4.4.1 Model comparison

Our research questions are: 'Which social media platform (Twitter or Facebook) and which type of data (PPI, MGC and UGC) has the most predictive power in box office sales predictions?'. Table 4.5 summarizes the average performance and standard deviations of the 6 models proposed in Section 4.3.2 in terms of RMSE, MAE, MAPE and R^2 across all 7 algorithms. We note that this table is based on the median 5x2cv results for each performance measure. We refer the reader to Appendix B for an overview of the median 5x2cv results in terms of RMSE, MAE, MAPE and R^2 for each model and performance algorithm. A first observation is the superiority of Facebook data in comparison to Twitter. From Table 4.5 it is clear that Facebook data has the best performance both with and without UGC

Table 4.5: Average (standard deviation) 5x2cv median RMSE, MAE, MAPE and R^2 across all algorithms

	Fb:base	Fb:plus	Tw:base	Tw:plus	FbTw:base	FbTw:plus
RMSE	1.7045	1.6123	2.0163	1.9226	1.6697	1.6333
	(0.4626)	(0.2840)	(0.3969)	(0.3897)	(0.4346)	(0.3109)
MAE	1.2961	1.2381	1.5537	1.4921	1.2826	1.2242
	(0.3209)	(0.2195)	(0.2980)	(0.2772)	(0.3189)	(0.1802)
MAPE	0.0861	0.0824	0.1043	0.1000	0.0857	0.0811
	(0.0221)	(0.0154)	(0.0197)	(0.0174)	(0.0237)	(0.0126)
R^2	0.4893	0.5038	0.2894	0.3425	0.5004	0.5315
	(0.2059)	(0.2053)	(0.1700)	(0.1509)	(0.2152)	(0.1616)

across all performance measures (i.e., Fb:base is superior to Tw:base and Fb:plus is superior to Tw:plus). Moreover, Facebook data is always superior to Twitter data even when comparing Facebook data without UGC (Fb:base) to Twitter data with UGC (Tw:plus). For example, Facebook outperforms the best Twitter model in terms of RMSE by at least 11%[6] and 16%[7] maximum, by 13% and 17% in terms of MAE, by 14% and 18% in terms of MAPE, and by 43%[8] and 47% [9] in the case of R^2. A second observation is that the inclusion of UGC always leads to better predictive performance for Facebook, Twitter, and the combination of both Twitter and Facebook. A final observation is that the combination of Facebook and Twitter data mostly leads to the best performance, except in terms of RMSE where Fb:plus is the best performer. However, we note that the differences between the models including only Facebook data and the models including both Facebook and Twitter data are rather small. Therefore, we take a look at the wins-ties-losses to test whether these observations are significant across the majority of algorithms.

Table 4.6 summarizes the wins-ties-losses across all 7 algorithms for each performance measure. We note that these counts are based on the 5x2cv median results for each performance measure in Appendix B. For example, the comparison of Fb:base against Fb:plus in terms of R^2 informs us that Fb:base wins in 3 out of the 7 times from Fb:plus and loses in 4 out of the 7 times in absolute numbers. However, when looking at the significant difference we notice that Fb:base and Fb:plus are tied in 6 cases and Fb:base both wins and loses in 1 case. Based on these findings we can draw several conclusions. First, the superiority of Face-

[6]This figure is calculated by comparing the performance of Fb:base and Tw:plus: $1 - (1.7045/1.9226) = 0.1134$.

[7]This figure is calculated by comparing the performance of Fb:plus and Tw:plus: $1 - (1.6123/1.9226) = 0.1614$.

[8]This figure is calculated as the increase in performance between Fb:base and Tw:plus: $((0.4893 - 0.3425)/0.3425) = 0.4286$.

[9]This figure is calculated as the increase in performance between Fb:plus and Tw:plus: $((0.5038 - 0.3425)/0.3425) = 0.4709$.

book over Twitter data is confirmed for models with and without UGC. We see that both in absolute and significant counts Fb:base and Fb:plus win in the majority of the cases for each performance measure. Hence, we can state regarding research question 1 that Facebook data holds the most predictive power. Second, we cannot confirm that the inclusion of UGC leads to a significant increase in predictive performance. For Twitter, Facebook; and a combination of Facebook and Twitter data the results are not significantly different with or without UGC in a majority of the cases. In most cases there is a significant tie between models with or without UGC. Finally, the combination of Facebook and Twitter data does not lead to a significant increase in performance when compared to models including Facebook data only. In most cases the comparison between Facebook and a combination of Facebook and Twitter leads to significant ties.

In summary, we can conclude that (1) Facebook data are superior to Twitter data, (2) the inclusion of UGC does not lead to a significant improvement, and (3) the combination of Twitter and Facebook does not significantly increase predictive performance when compared to solely Facebook data.

4.4.2 Algorithm performance

A secondary question we want to solve is: 'Which algorithm performs best in predicting box office sales?'. In Table 4.7 we summarize the averages and standard deviations across all models for each algorithm in terms of RMSE, MAE, MAPE and R^2. For detailed 5x2cv median results for each performance measure, we again refer the reader to Appendix B. The RMSE ranges from 1.4644 to 2.3994, the MAE from 1.1133 to 1.7723, the MAPE from 0.0746 to 0.1181, and the R^2 from 0.1424 to 0.5870. We see that RF is the best algorithm in terms of RMSE and R^2, followed by BT, GBM, KNN, DT, LR and NN. In terms of MAE and MAPE, BT comes out as the top performer. These results are partially in line with Cui et al. [53] who find that non-linear ensemble techniques such as RF and GBM perform best in forecasting future sales.

Table 4.8 provides the average ranks across all models in terms of RMSE, MAE, MAPE and R^2 with the results of the Friedman test. The Friedman test informs us that we can reject the null hypothesis that there are no significant differences between the algorithms for RMSE ($\chi^2(6) = 33.50$), MAE ($\chi^2(6) = 32.50$), MAPE ($\chi^2(6) = 32.35$) and R^2 ($\chi^2(6) = 32.57$). Hence, we perform the Bonferroni-Dunn post-hoc test to see which algorithms are significantly different from each other (i.e., difference bigger than the critical difference of 3.290485). In Table 4.8 the best performing algorithm is highlighted in bold and underlined. The algorithms for which the critical difference is smaller than 3.290485 (i.e., not significantly different) are expressed in bold font. Based on the average ranks RF is found to the best algorithm for all performance measures. Only in terms of MAE,

Table 4.6: Absolute (significant) wins-ties-losses across all 7 algorithms in terms of RMSE, MAE, MAPE and R^2

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

Table 4.7: Average (standard deviation) performance across all models based on RMSE, MAE, MAPE and R^2

	LR	RF	GBM	NN	KNN	DT	BT
RMSE	2.1382	1.4644	1.5507	2.3994	1.5724	1.7230	1.4704
	(0.1145)	(0.1778)	(0.1668)	(0.3698)	(0.2001)	(0.1867)	(0.1300)
MAE	1.6496	1.1295	1.2243	1.7723	1.2538	1.2917	1.1133
	(0.1057)	(0.1583)	(0.1364)	(0.2981)	(0.1713)	(0.1159)	(0.0982)
MAPE	0.1121	0.0755	0.0814	0.1181	0.0812	0.0864	0.0746
	(0.0083)	(0.0098)	(0.0104)	(0.0188)	(0.0138)	(0.0100)	(0.0065)
R^2	0.1424	0.5870	0.5291	0.2456	0.5559	0.4553	0.5845
	(0.0811)	(0.1134)	(0.1118)	(0.1606)	(0.1237)	(0.1060)	(0.0920)

Note: LR stands for linear regression, RF for random forest, GBM for gradient boosting machines, NN for neural networks, KNN for k-nearest neighbors, DT for decision trees and BT for bagged trees

Table 4.8: Average ranks across all models based on RMSE, MAE, MAPE and R^2 with critical difference 3.290485

	LR	RF	GBM	NN	KNN	DT	BT	Friedman χ^2 (6)
RMSE	6.33	**1.50**	**3.17**	6.67	**3.67**	5.00	**1.67**	33.50, p<0.001
MAE	6.33	**1.50**	**3.67**	6.67	**4.00**	**4.33**	**1.50**	32.50, p<0.001
MAPE	6.33	**1.33**	**3.50**	6.67	**3.67**	**4.67**	**1.83**	32.35, p<0.001
R^2	6.67	**1.33**	**3.50**	6.33	**3.17**	5.00	**2.00**	32.57, p<0.001

RF is tied with BT. For RMSE and R^2 BT, GBM and KNN are not significantly different from RF. For MAE and MAPE, DT is also not different in statistical terms. NN is the worst performing algorithm in terms of RMSE, MAE and MAPE and LR in terms of R^2. Readers who are interested in the average ranks together with the Friedman test with Bonferroni-Dunn post-hoc test for each performance measure separately are referred to Appendix C.

4.4.3 Information-fusion sensitivity analysis

Our final research question was: 'Which variables are most important?'. More specifically we are interested in which variables from which platform and which data type 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. 4.10 are calculated as the 5x2cv median mean increase in RMSE of permuting variable j in algorithm i, whereas the weights (w_i) are the weighted averages of 5x2cv median RMSEs of the FbTw:plus model (see final column Appendix B.1). To determine which variables are important, we made a plot that depicts the top 100 predictors against their sensitivity

score in decreasing order (black solid line in Figure 4.2). This means that the variable with the highest sensitivity score receives rank 1, the second highest sensitivity score rank 2 and so on. We also included the cumulative percentage of the sensitivity scores (red dotted line in Figure 4.2) in order to conduct a pareto-analysis [174]. The pareto rule states that with 20% of the variables 80% of the cumulative predictive performance can be achieved on average. However, these results can slightly differ depending on the data and the prediction algorithms. We note in Figure 4.2 that the cutoff to achieve 80% of the cumulative predictive performance corresponds with 23 variables (or $23/210 = 10.95\%$). Table 4.9 summarizes the top 23 variables based on information-fusion sensitivity analysis. We refer the reader to Appendix A for a description of the variables. Next to the rank, the variable and the sensitivity score we also added a column specifying the platform (i.e., FB or TW), the data type (i.e., PPI, MGC or UGC), the type of WOM (i.e., volume (vol), valence (val) or a combination of both (comb)) and whether or not the variable is time-restricted (T) or time-unrestricted (U).

First, the results indicate that Facebook is the most important social media platform with 83% of the most important variables related to FB and only 17% related to TW. This finding confirms the results of Oh et al. [173], namely that Twitter data become insignificant when including Facebook data. Second, most of the top predictors are related to UGC (83%), followed by PPI (13%) and MGC (4%). Hence, we confirm the findings of Goh et al. [101] that UGC has more impact on firm performance than MGC. Another important result in that PPI variables are the top predictors of box office revenues, despite the fact that their numbers are less extensive than UGC. For example, the total number of Facebook Page likes is the most important variable, the number of Talking About on Facebook is ranked fifth and the total number of Twitter followers is at rank eight. This can be explained by consumer engagement behavior (CEB) theory as follows [41]. Personal engagement, expressed as liking the Facebook page er following the Twitter page, reflects intrinsic motivation to go and see the movie and is highly correlated with box office sales. Interactive engagement, expressed as Facebook Talking About, represents engagement in the community and is also of major importance. Next to PPI, UGC is also a very important driver of box office sales. When we look at WOM variables in particular, we find that volume measures are most important (67%), followed by a combination of volume and valence (24%) and valence (9%). For example, the hype factor of the comments on Facebook is the second most important variable. The closer this variable is to 1, the higher the number of unique users who commented on Facebook, and hence the higher the amount of WOM, and the higher the box office sales [97]. Hence, the awareness effect (volume WOM) is more important than the persuasive effect (valence WOM). The most popular and important way of generating WOM on Facebook is by means of commenting. Finally, we found that both time-restricted (43%) and unrestricted

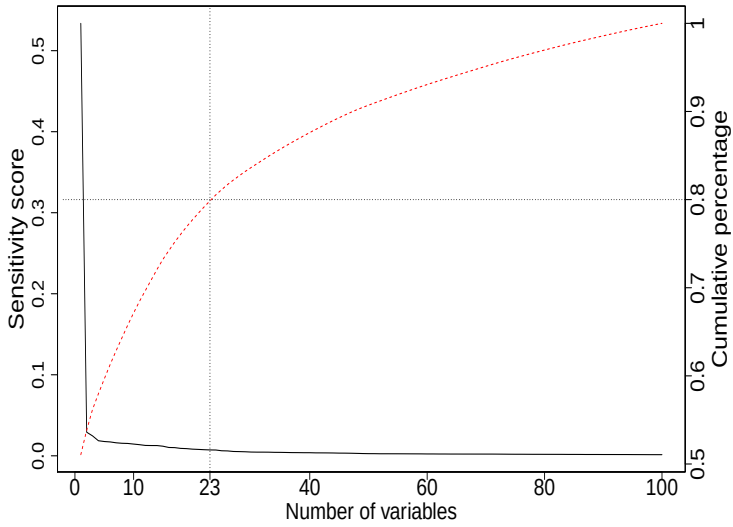


Figure 4.2: Scree and pareto plot of the cumulative sensitivity scores of the top 100 variables

(57%) UGC and MGC measures were important. However, time-based measures after the release of the movie were among the most important variables.

Table 4.9: Top 23 variables based on information-fusion sensitivity analysis

Rank	Variable	Sensitivity score	Platform	Data	WOM	U/T
1	NrOfLikesFb	0.5337	FB	PPI	-	-
2	HypeFactorCommFb	0.0291	FB	UGC	Vol	U
3	NrNeutral2WeeksAftrRelCommFb	0.0246	FB	UGC	Comb	T
4	NrTalkingAboutFb	0.0186	FB	PPI	-	-
5	NrPosCommFb	0.0176	FB	UGC	Comb	U
6	PctAftrRelCommFb	0.0171	FB	UGC	Vol	T
7	NrPosAftrRelCommFb	0.0161	FB	UGC	Comb	T
8	NrOfFollowersTw	0.0155	TW	PPI	-	-
9	AvgNrOfFavoritedTw	0.0152	TW	UGC	Vol	U
10	NrBfrRelCommFb	0.0145	FB	UGC	Vol	T
11	AvgNrOfCommPerDay2WeeksAftrRelFb	0.0137	FB	UGC	Vol	T
12	AvgNrOfRetweetedRepliesTw	0.0128	TW	UGC	Vol	U
13	AvgNrOfDaysBetweenCommFb	0.0127	FB	UGC	Vol	U
14	Nr2WeeksAftrRelCommFb	0.0126	FB	UGC	Vol	T
15	AvgLengthPostFb	0.0118	FB	MGC	Vol	U
16	RatioPosNegCommFb	0.0103	FB	UGC	Val	U
17	Pct2WeeksAftrReleaseFb	0.0101	FB	MGC	Vol	T
18	NrNeg2WeeksAftrRelCommFb	0.0091	FB	UGC	Comb	T
19	AvgNrOfRepliesTw	0.0089	TW	UGC	Vol	U
20	NrNegBfrRelCommFb	0.0082	FB	UGC	Comb	T
21	AvgNrOfCommFb	0.0079	FB	UGC	Vol	U
22	NrofCommFb	0.0075	FB	UGC	Vol	U
23	AvgSentScoreCommFb	0.0072	FB	UGC	Val	U

Note: FB represents Facebook, TW Twitter, PPI page popularity indicators, MGC marketer-generated content, UGC user-generated content, Val valence, Vol volume, Comb combination of volume and valence, T time-based, and U unbounded.

4.5 Discussion and implications

Our results provide important insights for researchers and practitioners wanting to study box office predictions. For practitioners, our findings provide an important framework to determine which data, algorithms and variables to include in the model. The choice of the best model for box office sales predictions is dependent upon two factors: (1) the available resources, and (2) whether or not the modeler is only interested in the predictions or he/she is also interested in the interpretability of the model. The available resources are seen as the available time to scrape, prepare and analyze the data and the computational execution time. If the available resources are low to moderate and the modeler is only interested in the predictions, our results indicate that the box office prediction model should only include Facebook data with PPI and MGC (i.e., the Fb:base model). We found in Section 4.4.1 that the difference in performance between models with UGC and without on Facebook is insignificant. However, if the available time and resources are high, we recommend to include Twitter and build a model with PPI and MGC from Facebook and Twitter (i.e., FbTw:plus). In case that the modeler is also interested in interpretability of the models, we recommend to always include UGC. Per Section 4.4.3, UGC variables are among the top drivers of box office sales. In terms of algorithms, our results suggest to use random forest. It is the top performing algorithm and it can be executed in parallel [19].

For researchers, our results provide important insights on both the methodological and the theoretical side. On the methodological side, our contributions are twofold. First, we give insight into which social media platform is most predictive of box office sales. Second, we benchmark several algorithms over different models and find that overall ensemble tree-based methods are superior (i.e., random forest and bagged trees). On the theoretical side, we contribute to literature by showing which variables from which data type are important and hence which marketing theories are the most important in explaining box office sales. We show that PPI and UGC are the most important drivers, whereas MGC is of lesser importance. This finding underscores the importance of personal and interactive engagement in explaining box office sales [173]. In terms of UGC and WOM, we conclude that volume is more important than valence. Hence, this implies that on social media the awareness effect is more important than the persuasive effect [151].

4.6 Conclusion and future research

In this study we assess (1) which social media platform performs best in predicting box office sales, (2) which algorithms performs best, and (3) which variables from which platform and which data type are driving the predictive performance. To

do so, 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 Facebook is more indicative of box office sales than Twitter in terms of RMSE, MAE, MAPE and R^2. We found that Facebook models outperformed the best Twitter model by at least 11% in RMSE, 13% in MAE, 14% in MAPE, and 43% in R^2. When comparing both platforms with and without UGC, Facebook models have significantly better performance than Twitter models in the majority of the cases. When assessing the added value of UGC over and above PPI and MGC, we did not find a significant improvement. When Facebook data or a combination of Facebook and Twitter data were used, there was an improvement on average, however this improvement was not significant for the majority of the algorithms. Finally, we found that the combination of Facebook and Twitter data did not significantly improve predictive performance beyond the Facebook data. In addition to comparing which social media platform and which data type performs best, we also assessed which algorithms performs best in predicting box office sales. The results show that RF was the top performer, followed by BT, GBM, KNN, DR, LR and NN for RMSE, MAPE and R^2. For MAE, RF and BT were tied on the first place. We note that the four top performing algorithms had equal performance in statistical terms. In general ensemble tree-based methods are the top performers.

Our information-fusion sensitivity analysis reveals that a PPI variable, the number of Facebook page likes, was the most important variables. The second most important variable is a UGC variable, the hype factor of the comments on Facebook. In general, UGC and PPI variables from Facebook dominated the top predictor list. In terms of WOM, we find that volume measures are more important than valence measures. More specifically, comments on Facebook were considered as the most important WOM indicator. These findings reveal the fact that personal engagement and interactive engagement mainly explain the influence of social media on box office sales, followed by the awareness and persuasive effect. In terms of time-restricted or unrestricted variables, unbounded variables are more numerous in the top predictor list but time-restricted variables are higher ranked on average.

Future research should focus on including more movies and making separate predictions for different types of movies. While we included a diverse set of movies, future studies can focus on movies with a specific budget range or a specific genre.

An interesting avenue for future research would be to include whether or not a post or a tweet was 'boosted'. Facebook (and also Twitter) strategically limit

the organic reach of posts to approximately 6.5% of the fan base, and event 2.5% for pages with more than 500,000 followers [158]. To increase their organic reach movie advertisers are forced to pay for additional reach. One can argue that by boosting a post (or a tweet) UGC volume increases and the importance of MGC decreases. As a result, UGC becomes more important than MGC within platform. To solve this issue we could detect which posts (and tweets) are boosted by detecting whether or not the number of likes of the particular post or tweet is bigger than 6.5% (or 2.5%) of the installed fan base. However, there are two major remarks to this approach. First, although the detection of boosting is possible in theory, this still does not change the fact that we do not have data about the advertising spent on Facebook and Twitter for each movie. Hence, we are still forced to work with a proxy and not with the true data. Second, the practice of boosting does not change our current conclusions or contributions. We want to investigate, given the current practice of boosting on each platforms, which platform, data types, algorithms and variables are the most important in predicting box office sales.

Another interesting 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 [68] or even movie success [141]. Since the motivation of this study was to compare Facebook and Twitter in box office sales, we chose the most general measure of box office sales (i.e., final box office gross sales).

A final suggestion for future research is to include information about the movie itself [141] in addition to social media data and the theaters' information (e.g., the number of visitors and the number screens) [130]. Moreover, one could also argue to include more social media platforms such as YouTube, Yahoo! Movies or Google trends data. These extensions are beyond the scope of this paper.

Finally, we would like to stress 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 movies, algorithms and variables. As a results, we believe this study makes a valuable contribution to literature on social media and box office sales from both the methodological and theoretical perspective.

4.7 Appendix

Appendix A: Predictors

A.1 Overview and explanation of Facebook variables (PPI= page popularity indicators, UGC = user-generated content, MGC = marketer-generated content, Vol = volume, Val = valence, Comb = combination volume and valence, U = unbounded, T = time-based)

Nr.	Variable name	Variable explanation	Data	WOM	U/T
1	NrOfLikesFb	Number of likes on Facebook page	PPI	-	-
2	NrOfPostsFb	Number of posts on Facebook page	MGC	Vol	U
3	NrTalkingAboutFb	Number of people talking about a Facebook page	MGC	Vol	U
4	AvgLengthPostFb	Average number of characters of a post	MGC	Vol	U
5	AvgNrOfDaysBetweenPostsFb	Average number of days between 2 posts	MGC	Vol	T
6	NrBeforeReleaseFb	Number of posts before the release date	MGC	Vol	T
7	NrAFterReleaseFb	Number of posts after the release date	MGC	Vol	T
8	PctBeforeReleaseFb	Percentage of the posts before the release date	MGC	Vol	T
9	PctAfterReleaseFb	Percentage of the posts after the release date	MGC	Vol	T
10	Nr1WeekBfrRelFb	Number of posts posted from 1 week before the release date until the release date	MGC	Vol	T
11	Pct1WeekBfrReleaseFb	Percentage of posts posted from 1 week before the release date until the release date	MGC	Vol	T
12	Nr2WeekAftrRelFb	Number of posts posted from the release date until 2 weeks after the release date	MGC	Vol	T
13	Pct2WeeksAftrReleaseFb	Percentage of posts posted from the release date until 2 weeks after the release date	MGC	Vol	T
14	NrPosFb	Number of positive posts	MGC	Comb	U
15	NrNegFb	Number of negative posts	MGC	Comb	U
16	NrNeutralFb	Number of neutral posts	MGC	Comb	U
17	RatioPosNegFb	Ratio positive versus negative posts	MGC	Val	U
18	PctPosFb	Percentage of posts being positive	MGC	Val	U
19	PctNegFb	Percentage of posts being negative	MGC	Val	U
20	PctNeutralFb	Percentage of posts being neutral	MGC	Val	U
21	NrPosBfrRelFb	Number of positive posts before the release date	MGC	Comb	T
22	NrNegBfrRelFb	Number of negative posts before the release date	MGC	Comb	T
23	NrNeutralBfrRelFb	Number of neutral posts before the release date	MGC	Comb	T
24	RatioPosNegBfrRelFb	Ratio positive versus negative posts before the release date	MGC	Val	T
25	PctPosBfrRelFb	Percentage of posts before release that are positive	MGC	Val	T
26	PctNegBfrRelFb	Percentage of posts before release that are negative	MGC	Val	T
27	PctNeutralBfrRelFb	Percentage of posts before release that are neutral	MGC	Val	T

28	NrPosAftrRelFb	Number of positive posts after the release date	MGC	Comb	T
29	NrNegAftrRelFb	Number of negative posts after the release date	MGC	Comb	T
30	NrNeutralAftrRelFb	Number of neutral posts after the release date	MGC	Comb	T
31	RatioPosNegAftrRelFb	Ratio of positive versus negative posts after the release date	MGC	Val	T
32	PctPosAftrRelFb	Percentage of posts after the release date that are positive	MGC	Val	T
33	PctNegAftrRelFb	Percentage of posts after the release date that are negative	MGC	Val	T
34	PctNeutralAftrRelFb	Percentage of posts after the release date that are neutral	MGC	Val	T
35	NrPos1WeekBfrRelFb	Number of positive posts from 1 week before the release date until the release date	MGC	Comb	T
36	NrNeg1WeekBfrRelFb	Number of negative posts from 1 week before the release date until the release date	MGC	Comb	T
37	NrNeutral1WeekBfrRelFb	Number of neutral posts from 1 week before the release date until the release date	MGC	Comb	T
38	NrPos2WeeksAftrRelFb	Number of positive posts from the release date until 2 weeks after the release date	MGC	Comb	T
39	NrNeg2WeeksAftrRelFb	Number of negative posts from the release date until 2 weeks after the release date	MGC	Comb	T
40	NrNeutral2WeeksAftrRelFb	Number of neutral posts from the release date until 2 weeks after the release date	MGC	Comb	T
41	RatioPosNeg2WeeksAftrRelFb	Ratio positive versus negative posts from the release date until 2 weeks after the release date	MGC	Val	T
42	PctPos2WeeksAftrRelFb	Percentage of posts from the release date until 2 weeks after the release date that are positive	MGC	Val	T
43	PctNeg2WeeksAftrRelFb	Percentage of posts from the release date until 2 weeks after the release date that are negative	MGC	Val	T
44	PctNeutral2WeeksAftrRelFb	Percentage of posts from the release date until 2 weeks after the release date that are neutral	MGC	Val	T
45	AvgSentScoreFb	Average sentiment score of the posts	MGC	Val	U
46	AvgSentBfrRelFb	Average sentiment score of posts before the release date	MGC	Val	T
47	AvgSentAftrRelFb	Average sentiment score of posts after the release date	MGC	Val	T
48	AvgSent2WeeksAftrRelFb	Average sentiment score of posts posted from the release date until 2 weeks after the release date	MGC	Val	T
49	ChangeSentPrePostRelFb	Change in sentiment score of the posts before and after the release date	MGC	Val	T
50	PctChangeSentPrePostRelFb	Percentage change in sentiment score of the posts before and after the release date	MGC	Val	T
51	AvgNrOfPostsPerDayPostsFb	Average number of posts per day	MGC	Vol	T

52	AvgNrOfPostsPerDayBfrRelPostsFb	Average number of posts per day before the release date	MGC	Vol	T
53	AvgNrOfPostsPerDayAftrRelPostsFb	Average number of posts per day after the release date	MGC	Vol	T
54	AvgNrOfPostsPerDay1WeekBfrRelPostsFb	Average number of posts per day 1 week before the release date	MGC	Vol	T
55	AvgNrOfPostsPerDay2WeeksAftrRelPostsFb	Average number of posts per day from the release date until 2 weeks after the release date	MGC	Vol	T
56	AvgNrOfLikesFb	Average number of likes on a post	UGC	Vol	U
57	NrofCommentsFb	Number of comments	UGC	Vol	U
58	AvgNrLikesCommentFb	Average number of likes on the comments	UGC	Vol	U
59	AvgNrOfCommentsFb	Average number of comments per post	UGC	Vol	U
60	AvgLengthCommentFb	Average number of characters per comment	UGC	Vol	U
61	AvgNrOfDaysBetweenCommentsFb	Average number of days between 2 comments	UGC	Vol	T
62	Nr1WeekBfrRelCommentsFb	Number of comments posted from 1 week before the release date until the release date	UGC	Vol	T
63	Pct1WeekBfrRelCommentsFb	Percentage of comments posted from 1 week before the release date until the release date	UGC	Vol	T
64	Nr2WeeksAftrRelCommentsFb	Number of comments posted from the release date until 2 weeks after the release date	UGC	Vol	T
65	Pct2WeeksAftrRelCommentsFb	Percentage of comments posted from the release date until 2 weeks after the release date	UGC	Vol	T
66	NrPosCommentsFb	Number of positive comments	UGC	Comb	U
67	NrNegCommentsFb	Number of negative comments	UGC	Comb	U
68	NrNeutralCommentsFb	Number of neutral comments	UGC	Comb	U
69	RatioPosNegCommentsFb	Ratio of positive versus negative comments	UGC	Val	U
70	PctPosCommentsFb	Percentage of comments that are positive	UGC	Val	U
71	PctNegCommentsFb	Percentage of comments that are negative	UGC	Val	U
72	PctNeutralCommentsFb	Percentage of comments that are neutral	UGC	Val	U
73	NrPosBfrRelCommentsFb	Number of positive comments before the release date	UGC	Comb	T
74	NrNegBfrRelCommentsFb	Number of negative comments before the release date	UGC	Comb	T
75	NrNeutralBfrRelCommentsFb	Number of neutral comments before the release date	UGC	Comb	T
76	PctPosBfrRelCommentsFb	Percentage of comments before release date that are positive	UGC	Val	T
77	PctNegBrfRelCommentsFb	Percentage of comments before release date that are negative	UGC	Val	T
78	PctNeutralBfrRelCommentsFb	Percentage of comments before release date that are neutral	UGC	Val	T
79	NrPosAftrRelCommentsFb	Number of comments after the release date that are positive	UGC	Comb	T

80	NrNegAftrRelCommentsFb	Number of comments after the release date that are negative	UGC	Comb	T
81	NrNeutralAftrRelCommentsFb	Number of comments after the release date that are neutral	UGC	Comb	T
82	RatioPosNegAftrRelCommentsFb	Ratio of positive versus negative comments after the release	UGC	Val	T
83	PctPosAftrRelCommentsFb	Percentage of comments after the release date that are positive	UGC	Val	T
84	PctNegAftrRelCommentsFb	Percentage of comments after the release date that are negative	UGC	Val	T
85	PctNeutralAftrRelCommentsFb	Percentage of comments after the release date that are neutral	UGC	Val	T
86	NrPos1WeekBfrRelCommentsFb	Number of positive comments from 1 week before the release until the release date	UGC	Comb	T
87	NrNeg1WeekBfrRelCommentsFb	Number of negative comments from 1 week before the release date until the release date	UGC	Comb	T
88	NrNeutral1WeekBfrRelCommentsFb	Number of neutral comments from 1 week before the release date until the release date	UGC	Comb	T
89	NrPos2WeeksAftrRelCommentsFb	Number of positive comments from the release date until 2 weeks after the release date	UGC	Comb	T
90	NrNeg2WeeksAftrRelCommentsFb	Number of negative comments from the release date until 2 weeks after the release date	UGC	Comb	T
91	NrNeutral2WeeksAftrRelCommentsFb	Number of neutral comments from the release date until 2 weeks after the release date	UGC	Comb	T
92	HypeFactorCommentsFb	Hype factor of the comments	UGC	Vol	U
93	NrBfrRelCommentsFb	Number of comments before the release date	UGC	Vol	T
94	NrAftrRelCommentsFb	Number of comments after the release date	UGC	Vol	T
95	PctBfrRelCommentsFb	Percentage of comments before the release date	UGC	Vol	T
96	PctAftrRelCommentsFb	Percentage of comments after the release date	UGC	Vol	T
97	AvgSentScoreCommentsFb	Average sentiment score of the comments	UGC	Val	U
98	AvgSentAftrRelCommentsFb	Average sentiment score of the comments after the release	UGC	Val	T
99	AvgNrOfCommentsPerDayCommentsFb	Average number of comments per day	UGC	Vol	T
100	AvgNrOfCommentsPerDayBfrRelCommentsFb	Average number of comments per day before the release date	UGC	Vol	T
101	AvgNrOfCommentsPerDayAftrRelCommentsFb	Average number of comments per day after the release date	UGC	Vol	T
102	AvgNrOfCommentsPerDay1WeekBfrRelCommentsFb	Average number of comments from 1 week before the release date until the release date	UGC	Vol	T

103	AvgNrOfCommentsPerDay2WeeksAftrRelCommentsFb	Average number of comments from the release date until 2 weeks after the release date	UGC	Vol	T

A.2 Overview and explanation of Twitter variables (PPI= page popularity indicators, UGC = user-generated content, MGC = marketer-generated content, Vol = volume, Val = valence, Comb = combination volume and valence, U = unbounded, T = time-based)

Nr.	Variable name	Variable explanation	Data	WOM	U/T
1	NrOfFollowersTw	Number of followers on a Twitter page	PPI	-	-
2	NrOfTweets	Number of tweets on a Twitter page	MGC	Vol	U
3	NrOfLikesTw	Number of likes on a Twitter page	PPI	-	-
4	AvgLengthTweet	Average number of characters of a tweet	MGC	Vol	U
5	AvgNrofDaysBetweenTweets	Average number of days between 2 tweets	MGC	Vol	T
6	NrBeforeReleaseTw	Number of tweets before the release date	MGC	Vol	T
7	NrAfterReleaseTw	Number of tweets after the release date	MGC	Vol	T
8	PctBeforeReleaseTw	Percentage of tweets before the release date	MGC	Vol	T
9	PctAfterReleaseTw	Percentage of tweets after the release date	MGC	Vol	T
10	Nr1weekBfrRelTw	Number of tweets posted 1 week before the release date until the release date	MGC	Vol	T
11	Pct1WeekBfrReleaseTw	Percentage of tweets posted from 1 week before the release date until the release date	MGC	Vol	T
12	Nr2WeekAftrRelTw	Number of tweets posted from the release date until 2 weeks after the release date	MGC	Vol	T
13	Pct2WeekAftrRelTw	Percentage of tweets posted from the release date until 2 weeks after the release date	MGC	Vol	T
14	NrPosTw	Number of tweets that are positive	MGC	Comb	U
15	NrNegTw	Number of tweets that are negative	MGC	Comb	U
16	NrNeutralTw	Number of tweets that are neutral	MGC	Comb	U
17	RatioPosNegTw	Ratio positive versus negative tweets	MGC	Val	U
18	PctPosTw	Percentage of tweets that are positive	MGC	Val	U
19	PctNegTw	Percentage of tweets that are negative	MGC	Val	U
20	PctNeutralTw	Percentage of tweets that are neutral	MGC	Val	U
21	NrPosBfrRelTw	Number of positive tweets before the release date	MGC	Comb	T
22	NrNegBfrRelTw	Number of negative tweets before the release date	MGC	Comb	T
23	NrNeutralBfrRelTw	Number of neutral tweets before the release date	MGC	Comb	T
24	RatioPosNegBfrRelTw	Ratio positive versus negative tweets before the release date	MGC	Val	T
25	PctPosBfrRelTw	Percentage of tweets before the release date that are positive	MGC	Val	T
26	PctNegBfrRelTw	Percentage of tweets before the release date that are negative	MGC	Val	T
27	PctNeutralBfrRelTw	Percentage of tweets before the release date that are neutral	MGC	Val	T

28	NrPosAftrRelTw	Number of positive tweets after the release date	MGC	Comb	T
29	NrNegAftrRelTw	Number of negative tweets after the release date	MGC	Comb	T
30	NrNeutralAftrRelTw	Number of neutral tweets after the release date	MGC	Comb	T
31	RatioPosNegAftrRelTw	Ratio positive and negative tweets after the release date	MGC	Val	T
32	PctPosAftrRelTw	Percentage of tweets after the release date that are positive	MGC	Val	T
33	PctNegAftrRelTw	Percentage of tweets after the release date that are negative	MGC	Val	T
34	PctNeutralAftrRelTw	Percentage of tweets after the release date that are neutral	MGC	Val	T
35	NrPos1WeekBfrRelTw	Number of positive tweets from 1 week before the release date until the release date	MGC	Comb	T
36	NrNeg1WeekBfrRelTw	Number of negative tweets from 1 week before the release until the release date	MGC	Comb	T
37	NrNeutral1WeekBfrRelTw	Number of neutral tweets from 1 week before the release date until the release date	MGC	Comb	T
38	NrPos2WeeksAftrRelTw	Number of positive tweets from the release date until 2 weeks after the release date	MGC	Comb	T
39	NrNeg2WeeksAftrRelTw	Number of negative tweets from the release date until 2 weeks after the release date	MGC	Comb	T
40	RatioPosNeg2WeeksAftrRelTw	Ratio positive versus negative tweets from the release date until 2 weeks after the release date	MGC	Val	T
41	PctPos2WeeksAftrRelTw	Percentage of tweets from the release date until 2 weeks after the release date that are positive	MGC	Val	T
42	PctNeg2WeeksAftrRelTw	Percentage of tweets from the release date until 2 weeks after the release date that are negative	MGC	Val	T
43	AvgSentScoreTw	Average sentiment score tweets	MGC	Val	U
44	AvgSentBfrRelTw	Average sentiment score tweets before the release date	MGC	Val	T
45	AvgSentAftrRelTw	Average sentiment score tweets after the release date	MGC	Val	T
46	AvgSent2WeeksAftrRelTw	Average sentiment score tweets from the release until 2 weeks after the release date	MGC	Val	T
47	ChangeSentPrePostRelTw	Change in sentiment score tweets before and after the release date	MGC	Val	T
48	PctChangeSentPrePostRelTw	Percentage change in sentiment score tweets before and after the release date	MGC	Val	T
49	AvgNrOfTweetsPerDayTweetsTw	Average number of tweets per day	MGC	Vol	T
50	AvgNrOfTweetsPerDayBfrRelTweetsTw	Average number of tweets per day before the release date	MGC	Vol	T
51	AvgNrOfTweetsPerDayAftrRelTweetsTw	Average number of tweets per day after the release date	MGC	Vol	T

52	AvgNrOfTweetsPerDay1WeekBfrRelTweetsTw	Average number of tweets per day from 1 week before the release date until the release date	MGC	Vol	T
53	AvgNrOfTweetsPerDay2WeeksAftrRelTweetsTw	Average number of tweets per day from the release date until 2 weeks after the release date	MGC	Vol	T
54	NrNeutral2WeeksAftrRelTw	Number of neutral tweets from the release date until 2 weeks after the release date	MGC	Val	T
55	PctNeutral2WeeksAftrRelTw	Percentage of tweets from the release date until 2 weeks after the release date that are neutral	MGC	Val	T
56	AvgNrOfFavoritedTw	Average number of times a tweet is favorited	UGC	Vol	U
57	AvgNrOfRetweetedTw	Average number of times a tweet is retweeted	UGC	Vol	U
58	PctIsRetweetTw	Percentage of tweets that are a retweet	UGC	Vol	U
59	NrofRepliesTw	Number of replies	UGC	Vol	U
60	AvgNrOfRepliesTw	Average number of replies per tweet	UGC	Vol	U
61	AvgNrOfRetweetedRepliesTw	Average number of times a reply is retweeted	UGC	Vol	U
62	AvgNrOfFavoritedRepliesTw	Average number of times a reply is favorited	UGC	Vol	U
63	AvgLengthReplyTw	Average number of characters per reply	UGC	Vol	U
64	AvgNrOfDaysBetweenRepliesTw	Average number of days between 2 replies	UGC	Vol	T
65	NrBfrRelRepliesTw	Number of replies before the release date	UGC	Vol	T
66	NrAftrRelRepliesTw	Number of replies after the release date	UGC	Vol	T
67	PctBfrRelRepliesTw	Percentage of replies before the release date	UGC	Vol	T
68	PctAftrRelRepliesTw	Percentage of replies after the release date	UGC	Vol	T
69	Nr1WeekBfrRelRepliesTw	Number of replies from 1 week before the release date until the release date	UGC	Vol	T
70	Pct1WeekBfrRelRepliesTw	Percentage of replies from1 week before the release date until the release date	UGC	Vol	T
71	Nr2WeeksAftrRelRepliesTw	Number of replies from the release date until 2 weeks after the release date	UGC	Vol	T
72	Pct2WeeksAftrRelRepliesTw	Percentage of replies from the release date until 2 weeks after the release date	UGC	Vol	T
73	NrPosRepliesTw	Number of positive replies	UGC	Comb	U
74	NrNegRepliesTw	Number of negative replies	UGC	Comb	U
75	NrNeutralRepliesTw	Number of neutral replies	UGC	Comb	U
76	RatioPosNegRepliesTw	Ratio positive versus negative replies	UGC	Val	U
77	PctPosRepliesTw	Percentage of replies that are positive	UGC	Val	U
78	PctNegRepliesTw	Percentage of replies that are negative	UGC	Val	U
79	PctNeutralRepliesTw	Percentage of replies that are neutral	UGC	Val	U
80	NrPosBfrRelRepliesTw	Number of positive replies before the release date	UGC	Comb	T
81	NrNegBfrRelRepliesTw	Number of negative replies before the release date	UGC	Comb	T
82	NrNeutralBfrRelRepliesTw	Number of neutral replies before the release date	UGC	Comb	T

83	PctPosBfrRelRepliesTw	Percentage of replies before the release date that are positive	UGC	Val	T
84	PctNegBrfRelRepliesTw	Percentage of replies before the release date that are negative	UGC	Val	T
85	PctNeutralBfrRelRepliesTw	Percentage of replies before the release date that are neutral	UGC	Val	T
86	NrPosAftrRelRepliesTw	Number of positive replies after the release date	UGC	Comb	T
87	NrNegAftrRelRepliesTw	Number of negative replies after the release date	UGC	Comb	T
88	RatioPosNegAftrRelRepliesTw	Ratio positive versus negative replies after the release date	UGC	Val	T
89	PctPosAftrRelRepliesTw	Percentage of replies after the release date that are positive	UGC	Val	T
90	PctNegAftrRelRepliesTw	Percentage of replies after the release date that are negative	UGC	Val	T
91	NrPos1WeekBfrRelRepliesTw	Number of positive replies from 1 week before the release date until the release date	UGC	Comb	T
92	NrNeg1WeekBfrRelRepliesTw	Number of negative replies from 1 week before the release date until the release date	UGC	Comb	T
93	NrNeutral1WeekBfrRelRepliesTw	Number of neutral replies from 1 week before the release date until the release date	UGC	Comb	T
94	NrPos2WeeksAftrRelRepliesTw	Number of positive replies from the release date until 2 weeks after the release date	UGC	Comb	T
95	NrNeg2WeeksAftrRelRepliesTw	Number of negative replies from the release date until 2 weeks after the release date	UGC	Comb	T
96	NrNeutral2WeeksAftrRelRepliesTw	Number of neutral replies from the release date until 2 weeks after the release date	UGC	Comb	T
97	AvgSentScoreRepliesTw	Average sentiment score replies	UGC	Val	U
98	AvgSentAftrRelRepliesTw	Average sentiment score replies after the release date	UGC	Val	T
99	AvgNrOfRepliesPerDayRepliesTw	Average number of replies per day	UGC	Vol	T
100	AvgNrOfRepliesPerDayBfrRelRepliesTw	Average number of replies per day before the release date	UGC	Vol	T
101	AvgNrOfRepliesPerDayAftrRelRepliesTw	Average number of replies per day after the release date	UGC	Vol	T
102	AvgNrOfRepliesPerDay1WeekBfrRelRepliesTw	Average number of replies per day from 1 week before the release date until the release date	UGC	Vol	T
103	AvgNrOfRepliesPerDay2WeeksAftrRelRepliesTw	Average number of replies per day from the release date until 2 weeks after the release date	UGC	Vol	T
104	HypeFactorRepliesTw	Hype factor replies	UGC	Vol	U
105	NrNeutralAftrRelRepliesTw	Number of neutral replies after the release date	UGC	Val	T
106	PctNeutralAftrRelRepliesTw	Percentage of replies after the release that are neutral date	UGC	Val	T

Appendix B: Median performance

B.1 Median 5x2cv RMSE

Median RMSE	Fb:base	Fb:plus	Tw:base	Tw:plus	FbTw:base	FbTw:plus
LR	2.0892	2.0930	2.3490	2.0117	2.1616	2.1246
RF	1.3516	1.3686	1.7425	1.6350	1.3228	1.3659
GBM	1.5304	1.4231	1.8121	1.6845	1.3738	1.4804
NN	2.5917	1.8985	2.7639	2.7282	2.4081	2.0062
KNN	1.4365	1.4439	1.8439	1.8169	1.4421	1.4511
DT	1.5321	1.6597	1.9367	1.9768	1.5932	1.6395
BT	1.3999	1.3995	1.6660	1.6053	1.3859	1.3658

B.2 Median 5x2cv MAE

Median MAE	Fb:base	Fb:plus	Tw:base	Tw:plus	FbTw:base	FbTw:plus
LR	1.6206	1.6383	1.8270	1.6186	1.6885	1.5048
RF	1.0182	1.0370	1.3895	1.2652	1.0336	1.0337
GBM	1.1877	1.1189	1.4296	1.3504	1.0816	1.1777
NN	1.8697	1.4199	2.0939	2.0478	1.7922	1.4103
KNN	1.1468	1.1336	1.4733	1.4764	1.1395	1.1535
DT	1.1729	1.2545	1.4271	1.4430	1.1954	1.2570
BT	1.0569	1.0646	1.2352	1.2432	1.0476	1.0322

B.3 Median 5x2cv MAPE

Median MAPE	Fb:base	Fb:plus	Tw:base	Tw:plus	FbTw:base	FbTw:plus
LR	0.1135	0.1121	0.1226	0.1086	0.1176	0.0983
RF	0.0689	0.0702	0.0908	0.0850	0.0691	0.0690
GBM	0.0789	0.0751	0.0969	0.0914	0.0701	0.0762
NN	0.1223	0.0937	0.1398	0.1335	0.1221	0.0974
KNN	0.0733	0.0726	0.0994	0.0987	0.0709	0.0725
DT	0.0752	0.0822	0.0969	0.1004	0.0797	0.0843
BT	0.0707	0.0712	0.0840	0.0820	0.0703	0.0697

B.4 Median 5x2cv R^2

Median MAPE	Fb:base	Fb:plus	Tw:base	Tw:plus	FbTw:base	FbTw:plus
LR	0.2200	0.0758	0.0145	0.1658	0.1647	0.2134
RF	0.6453	0.6587	0.3949	0.4997	0.6683	0.6548
GBM	0.5809	0.5665	0.3636	0.4171	0.6423	0.6044
NN	0.1660	0.4440	0.0831	0.1157	0.2207	0.4442
KNN	0.6381	0.6395	0.3806	0.4132	0.6358	0.6286
DT	0.5485	0.4981	0.3301	0.3109	0.5323	0.5117
BT	0.6265	0.6442	0.4587	0.4753	0.6390	0.6636

Appendix C: Friedman test with Bonferroni-Dunn post-hoc test

C.1 RMSE (CD = 2.548799)

RMSE	LR	RF	GBM	KNN	DT	BT	NN	Friedman χ^2 (6)
Fb:base	6.2	2.2	4	2.6	4.3	1.9	6.8	47.96, p<0.001
Fb:plus	6.7	1.5	3.6	3	5.1	2	6.1	52.11, p<0.001
Tw:base	6.1	2.8	4	2.9	4.1	1.2	6.9	49.97, p<0.001
Tw:plus	5.8	1.7	3.1	4.1	4.8	1.7	6.8	49.54, p<0.001
FbTw:base	6.1	2	3	3.2	4.7	2.1	6.9	48.34, p<0.001
FbTw:plus	6.6	1.7	3.4	3.2	5	1.9	6.2	49.93, p<0.001

C.2 MAE (CD = 2.548799)

MAE	LR	RF	GBM	KNN	DT	BT	NN	Friedman χ^2 (6)
Fb:base	6.3	1.8	4.2	3.4	3.9	1.7	6.7	49.54, p<0.001
Fb:plus	6.6	1.5	3.8	3	4.9	2	6.2	50.79, p<0.001
Tw:base	6.1	2.4	4	4	3.6	1	6.9	52.59, p<0.001
Tw:plus	5.9	1.7	3.3	4.5	4.1	1.7	6.8	48.81, p<0.001
FbTw:base	6.2	1.9	3.2	3.6	4.6	1.7	6.8	50.44, p<0.001
FbTw:plus	6.3	1.5	3.4	3.6	5.1	1.8	6.3	50.14, p<0.001

C.3 MAPE (CD = 2.548799)

MAPE	LR	RF	GBM	KNN	DT	BT	NN	Friedman χ^2 (6)
Fb:base	6.3	1.8	4.2	3.4	3.9	1.7	6.7	49.54, p<0.001
Fb:plus	6.6	1.5	3.8	3	4.9	2	6.2	50.79, p<0.001
Tw:base	6.1	2.4	4	4	3.6	1	6.9	52.59, p<0.001
Tw:plus	5.9	1.7	3.3	4.5	4.1	1.7	6.8	48.81, p<0.001
FbTw:base	6.2	1.9	3.2	3.6	4.6	1.7	6.8	50.44, p<0.001
FbTw:plus	6.3	1.5	3.4	3.6	5.1	1.8	6.3	50.14, p<0.001

C.4 R^2 (CD = 2.548799)

R^2	LR	RF	GBM	KNN	DT	BT	NN	Friedman χ^2 (6)
Fb:base	6.4	2.2	4.2	1.9	4.5	2.2	6.6	50.79, p<0.001
Fb:plus	6.8	1.9	3.9	2	4.9	2.6	5.9	48.51, p<0.001
Tw:base	6.8	2.4	4.1	2.7	4.4	1.4	6.2	51.13, p<0.001
Tw:plus	6.1	1.6	3.1	3.4	4.8	2.5	6.5	43.89, p<0.001
FbTw:base	6.8	1.5	3.3	2.3	4.9	3	6.2	51.69, p<0.001
FbTw:plus	6.6	1.8	3.6	2.4	5.2	2.3	6.1	49.41, p<0.001

5 Conclusion

5.1 Discussion

In light of the recent scandals concerning the Facebook data breach and Cambridge Analytica, the question how companies use your social media data for targeted advertising becomes more relevant than ever. Since our approach shows much resemblance with the Cambridge Analytica case, we discuss the Cambridge Analytica case, the differences with our approach, and the implications on this dissertation more in detail.

Cambridge Analytica was founded in 2013 as an offshoot company of the SCL group [119]. The company originally positioned itself as a data mining and consumer research firm. Their customer base were mainly political and corporate clients. In the beginning of 2014 Cambridge Analytica acquired data of 87 million Facebook users by an application developed by an academic researcher. The application had over 270,000 participants, who gave their approval to gather their personal data. In addition, the application also gathered the personal data of the friends of the participants. The application itself consisted of several steps [112]. First, the application took a survey of 120 questions to discover the personality traits of the users. To fill out the survey the users were paid on average 5$. Next, at the end of the survey the users had to give permission to scrape there Facebook data to get paid. The application itself gathered all Facebook information of the users. Next to that, the application also gathered the data of all the friends of the participants giving a total amount of 87 million Facebook users in their database.

Cambridge Analytica then used the data from the 270,000 users who completed the personality quiz as a training set for their ensemble models. The independent variables in their case were the Facebook likes and the dependent variable was the personality of the user. Afterwards, Cambridge Analytica predicted the personality of the remaining 87 million people based on training model of the 270,000 users. With these personality predictions customized advertisements were created to target voters in the US 2016 elections.

It is clear from the previous paragraph that our data collection in Chapter 2 and Chaper 3 is more or less the same as the Cambridge Analytica case. We made a customized application that extracted the data of the participants of the application. We did this in cooperation with a European soccer team. In total, 5,000 users allowed us to gather their data. Next to that, our application also gathered the data of friends of the users. We then used the Facebook variables of these users and their friends to assess the predictive capacity of Facebook on the user and network level. Besides the data collection, there are several crucial differences between our approach and the Cambridge Analytica case. First, in the beginning of the application we explicitly mentioned the users that there Facebook data were being collected. Moreover, we also included a rules and regulations section that stated which data were being gathered to the users and that the data would solely be used for academic purposes. We also included our contact information for any further questions. Second, we did not include a personality test in our application, so we do not have information about the personality traits of the users. Third, by the time of the data collection in 2014, the Facebook rules and regulations (and also the privacy regulations) allowed that the data of the friends were gathered. This was also confirmed by the legal department of our university at that time. Finally, the information of the friends is less complete than the user information in our case. For example, we do not have data to which comments or photos the friends replied or liked.

All of the above differences make it clear that there is a very crucial difference between the Cambridge Analytica case and our case: Cambridge Analytica violated the rules and regulations of Facebook and the general privacy law at that time by selling their data to private companies. To make sure that our data would only be used for academic purposes and would not be abused by companies, we took the following steps. First, our agreement with the European soccer company stipulated that only researchers of Ghent University could have access to the Facebook data and that the data would never be shared with the European soccer team. Only the analyses and results were shared. Second, the data were stored on an external server were only a few people (i.e., researchers working on the data) had access to the data. Finally, when this dissertation and papers under revision are finished the data will be destroyed to make sure there is no abuse.

Given the Cambridge Analytica case, we believe that this dissertation provides

important insight to the general public on how advertisers can (mis)use their social media data. This dissertation identifies how to use social media data for targeted advertising purposes. We include a general framework on how companies can do this. Moreover, we also elaborate on the algorithms and the variables necessary to make these targeting models. Our list of important variables throughout the different chapters gives Facebook users on indication which data advertisers use for their campaigns. We hope that with this list (and this dissertation in general) we can make people more conscious about the information they put online and the dangers of social media usage. Also, we hope that this on its turn makes users more skeptical and critical towards practices such as Cambridge Analytica.

5.2 Conclusion and implications

In this dissertation we set out to harness the power of social media data in different applications and on different levels of analysis. Our goal was to provide evidence that social media data have predictive value and as such firms can implement a one-to-one advertising strategy on social media. To do so, we employed a data analytical strategy that assessed the predictive and descriptive capacity of social media data. In the predictive phase, our system estimated and compared several state-of-the-art prediction algorithms. In the descriptive phase, we assessed which variables were driving predictive performance. First, we discuss the general insights of this dissertation. Next, we summarize each chapter according to their methodological and theoretical contributions.

5.2.1 General findings

In this section we take a step back and discuss how this dissertation harnesses the predictive power of social media from different perspectives. Remember that the main questions of this dissertation were: (1) 'Is it feasible to use social media for predictive purposes?', (2) 'Which algorithms are most important?', and (3) 'Which variables are most important?'.

Regarding the first question the results indicate that it is feasible to use social media data as input data for predictive models. We found that this can be done with a high predictive accuracy. However, this accuracy is highly dependent upon the available data. Our findings indicate that the more individual user data is available, the higher the predictive accuracy. For example, in Chapter 2 we have a lot of socio-demographic and behavioral variables and achieve an AUC up to 80.38%. In Chapter 3 we find an AUC up to 97.59%. We notice that in Chapter 3 our level of analysis is the network level, but we have a lot of detailed information about the interactions between ego and alter on an individual level. For example, the number of photo tags between ego and alter, and the number of status comment

from ego to alter. In Chapter 4 ,on the other hand, we focus on the aggregate product performance level. In this case we do not have same level of granularity as in the previous studies. For example, we have information about the comments and likes users place on the official Facebook page, but we do not have information about the personal likes and interests of the users. Hence, we can only include user behavior that is observed on public Facebook pages on an aggregate level in our predictive models. These findings are also substantiated by Martens et al. [159] who conclude that the inclusion of fine-grained non-aggregate customer data leads to superior performance in predictive analytics.

Another interesting observation regarding research question 1 is that the methodologies to assess the predictive capacity of Facebook are highly dependent upon the business context. In Chapter 2 our data was provided by a European soccer team. Therefore we decided to focus on event attendance behavior of users that already attended a soccer game to avoid sample selection issues. In this case, our dependent variable (i.e., event attendance) was not highly skewed, so we did not had to control for class imbalance. However, in Chapter 3 we focused on social ties and were confronted with the long tail problem in social networks [75]. This problem states that users on social media only interact regularly with a small percentage of their total network. Since each user has one romantic tie in his/her network, this is translated into a high class imbalance. Hence, in this case it is necessary to control for class imbalance with data sampling techniques. Chapters 2 and 4 only include time and frequency variables, Chapter 4 also adds text and sentiment variables. Since we do not have a lot of individual user-level data available, text and sentiment analysis are necessary to extract the maximum amount of information from the user-generated content. For the marketer-generated content text and sentiment analysis are even more crucial since we only have text data available.

Regarding which algorithms perform best the following conclusion can be drawn. The results indicate that tree-based ensemble methods are superior in both classification and regression problems. However, there is no single best algorithm in data mining and as such the superiority of a certain algorithm is always dependent upon the characteristics of the data set and the assumptions of the algorithm. In that regard Wolpert's no free lunch theorem states that ,when comparing two algorithms A and B, there are just as many situations where A is superior to B and B is superior to A [219]. In the case of social media prediction the reasons of the superiority of tree-based ensembles (i.e., adaboost and random forest) are manifold. First, tree-based methods are non-parametric techniques that do not require the normality assumption to be met [19]. As in many real-life data sets, the analyses suggest that the normality assumption is not met. For example, the superior performance of tree-based methods over logistic and linear regression suggests that the data are not normally distributed and non-linear. A second reason is that ad-

aboost and random forest are ensemble methods. Ensemble methods lower the test set error by solving the computational, representational and statistical problems of single prediction algorithms [66]. Moreover, when confronted with a lot of variables, single prediction algorithms tend to be unstable and overfit [57]. Random forest reduces the variance of decision trees by combining bootstrap aggregation with random subspaces [35]. Stochastic boosting does not only decrease the variance of decision trees but also lowers the bias component [94]. This explains why adaboost and random forest are the overall top performers.

In terms of important variables we notice that variables that relate to the user's behavior are one of the top predictors. Depending on the business context, its ranking can differ. For example, in Chapter 2 the number of events attended in the past was one of the top predictors, whereas in Chapter 4 UGC was omnipresent in our top list of predictors. Next, we find evidence of the theory of homophily: 'birds of a feather flock together' [160]. This theory states that people who are alike often group together and share the same opinions and interests. In Chapter 2, we find that people have a higher chance of attending an event if their friends are attending as well. In Chapter 3 we observe a positive relationship between the number of common likes and interests and having a romantic partnership. Finally, we found that variables representing actions that require more effort (e.g., commenting) have a higher impact on predictive performance. This is explained by social signaling theory, which states that the more time people invest in a certain task, the stronger their relationship with the user or the product [202]. For example, in Chapter 3 we found that variables related to comments were the most important frequency and time-related variables in predicting romantic ties. In Chapter 4 we also found that comments are the most important communication type on Facebook in explaining box office sales.

Finally, when looking at the relationship between the top predictors and the response variable we observe that not all relationships have the expected directions. This is explained by the fact that other studies mostly rely on high level user data, whereas we include user data at its most granular level. In Chapter 2 for example, we find that the number of friends that attend the event have a positive effect on event attendance at first but afterwards a negative effect. However, from previous literature [5] showed that the adoption probability is higher when more friends adopt. In the case of Facebook events we find that this is only true for the first 12 that adopt. Another surprising relationship between was found between the recency of comments on status and photo tags and the probability of being a significant other in Chapter 3. In contrast to previous literature [9], we find that the longer it has been since the last comment on a post (or photo tag), the higher the propensity of being a significant other.

5.2.2 Contributions of each study

Chapter 2 assesses the predictive power of Facebook on the most granular level: the user. The study evaluates whether or not a user's friends data can improve the performance of event attendance prediction models over and above user data. Our methodological insights are that the Facebook friends data significantly improve model performance in a majority of the algorithms. These findings are also substantiated by the fact that the absolute and relative number of friends were among the top predictors of event attendance. At the theoretical side, we find evidence of homophily [160], social influence [89] and trust [120] in event attendance prediction. However, our findings indicate that only for the first (close) friends that indicate to attend the focal event the probability of attendance rises, afterwards the probability decreases once a threshold has been reached. This relationship can be explained by the fact that Facebook stops propagating the event through the News Feed when a lot of friends have indicated to attend the event to avoid spamming [81].

Chapter 3 focuses on the network level. This study uses disaggregated features to predict romantic ties between ego and alter. Disaggregated features are separate measures computed per interaction (i.e., commenting, liking, and tagging) and post type (i.e., statuses, photos, albums, videos, check-ins, and location updates). From a methodological perspective this study shows that we can predict romantic ties with high predictive power. From a theoretical side our findings extend the current theories on social tie prediction. We show that disaggregated features should be included in the model to uncover the true relationship between predictor and response in predicting romantic ties. For example, we find a positive relationship between the recency of a status comment and the probability of being a significant other. The reason is that after some time couples spend more time together and replace online with offline communication [40].

Chapter 4 investigates the impact of social media data on the most aggregate level (i.e., the product level). The goal of this study is to determine which social media platform (Facebook or Twitter) is most indicative of box office sales. On the methodological side this study is the first to conduct such a thorough analysis of the two social media platforms both in terms of algorithms and variables. The study finds that Facebook is significantly better in predicting box office sales than Twitter. Moreover, the study also shows that user-generated content does not significantly increase predictive performance. On the theoretical side this study contributes to literature by evaluating which content type is the driving force of predictive performance. The results indicate that consumer engagement behavior theory is one of the most important marketing theories in explaining box office sales [41]. When looking at solely word-of-mouth variables, our findings show that volume is more important than valence. Hence, this implies that the awareness effect dominates the persuasiveness effect in box office sales predictions [151].

5.3 Limitations and future research

5.3.1 General limitations

A first limitation of this dissertation is related to the positioning of the dissertation. In this dissertation we set out the harness the power of social media in several predictive analytics applications. However, in a strictu sense, a predictive analytics study involves the implementation of a time window which defines an independent (or calibration) period and a dependent (or prediction) period [17]. In that case there is a clear separation between the information that is available at a certain time and the future event. For example, to determine Facebook usage increase Ballings and Van den Poel [19] asked the users to run the Facebook application twice. The first run determined the independent variables. The second run recollected the same data from the same users to figure out whether the user increased his/her usage when compared to the first data dump. In our case we only collected the data at one point in time and did not conduct a second data collection in a subsequent time period with the same users. Hence, we only have cross-sectional data with aggregated variables until a certain time period, therefore we used an in-period test set instead of an out-of-period test set. For example, in Chapter 2, most of the predicted events happen after the time of our data collection (from May 7 to June 9, 2014), however some events occur during the time of our data collection. The same reasoning can be applied to Chapter 4 where gross box office revenues are explained using a large variety of variables ranging from unrestricted to time restricted variables. In a purely predictive context, only variables before the release date should be included to forecast box office sales. While this dissertation does not employ a pure predictive setting with a strictly defined time window, we still employ the same methodologies as in other predictive analytics studies. We use prediction (or in a broader sense data mining or machine learning) models to estimate the relationship between predictors and response. When confronted with unseen data, these models can be employed to predict future events. For example, if we would conduct a new data collection, we can supply our current models with this unseen data to make predictions. Finally, we note that in other fields such as health care analytics the term 'predictive analytics' often refers to the same setting as this dissertation [207].

A second limitation of this dissertation is closely related to the problem of endogeneity and reversed causality. In our current setting it is hard to know whether the predictors impact our outcome of interest or vice versa. For example, in Chapter 2 your event attendance could be influenced by the number of friends that are attending, but your event attendance itself could also impact the number of friends that attend the focal event. The same reasoning goes for Chapter 4: does intensified social media activity leads to more box office sales, or do higher box office sales intensify social media buzz? A first solution to this problem is related to the

first limitation, namely the use of a purely predictive setting with a time window. In that case all the predictors strictly happen before the response and you can confirm that precedent behavior predicts the future event. However, even in that case it is hard to establish true causality. Another way to address this issue is explained in Ballings et al. [23]. The best way of getting the ground truth of a causal relationship is to perform an intervention on one of the two variables and see whether this changes the distribution of the other variable [166]. However, for box office revenues or event attendance this is a hard to do because the extracted variables are beyond the control of the researcher. Another possibility is to look at the joint distribution between two variables and investigate whether there are asymmetries between cause and effect. This would enable us to find out the causal direction with a reasonable reliability (i.e., thus X causes Y or does Y causes X). To do so, we propose to use the approach of Hoyer et al. [117]. This approach assumes that the effect is a function of the cause and some additive independent noise. To test whether X causes Y one must regress Y on X, compute the residuals $(Y - f(x))$ for all observations and test if the residuals are independent of X. To regress Y on X, we can use a Gaussian Process Regression. To test the independence of the residuals and X, we can use the Hilbert Schmidt Independence criterion with gamma approximation and heuristically chosen kernel bandwidths. This null hypothesis assumes independence of X and computes the p-values. To determine whether X causes Y or Y causes X, we take the model with highest p-value for independence between residuals and X [166].

A third limitation is that this dissertation relies upon non-parametric statistical tests to discover significant differences between models and algorithms. The motivation for these statistical tests are based on the work of Demšar [64]. However, the non-parametric tests assume independence between the data sets and are only intended for comparing multiple classifiers across different data sets. In this dissertation we test these non-parametric test on multiple classifier runs over a cross-validation on the same data set. Demšar [64] notes that this implies that the datasets and the performance results are correlated, and this leads to an inflated type 1 error of these statistical tests. Hence, testing multiple algorithms classifiers on repetitions of the same data set requires other statistical tests. In literature, the comparison of algorithms across one data set has received much less attention. As to date there is no single best solution available. Burez and Van den Poel [39] discuss two possible methods when working with 5x2cv and multiple algorithms on one dataset. When comparing the error rate on two classifiers, they suggest to use the combined 5x2cv F-test as mentioned by Alpaydin [1]. This test is an improvement of the F-test proposed by Dietterich [65] and achieves a lower type 1 error rate and a higher power. If you are only testing for statistical differences in terms of AUC, several authors use the Delong test to compare two AUCs [63]. However, in the case of 5x2cv these AUCs are correlated (i.e., algorithms are ap-

plied to the same test each time). The authors therefore suggest that the averages of two algorithms are significantly different if on at least 7 out of the 10 runs the DeLong statistic was significant. For the sake of consistency, we suggest to use the combined 5x2cv Alpayadin F-test in other future applications. The p-values of this F-test can also be corrected for family-wise error as suggested by García et al. [98]. As a final note, we would like to add that although these non-parametric tests cannot be used in strictu sense to compare multiple classifiers on the same data set, these tests are often employed in practice (see for example Ballings and Van den Poel [19], Coussement and De Bock [50], and Coussement et al. [49]).

A fourth limitation is that we do not tune the hyper-parameters of all algorithms. For example, we use the default parameters for random forest, stochastic boosting, and bagging. For random forest it is most common to tune the number of random features to select at each tree split [184]. For stochastic boosting the number of iterations is mostly tuned to avoid overfitting [88]. In bagging the number of bags in the ensemble are often cross-validated [57]. Preliminary analyses of tuning these hyper-parameters across each study revealed that tuning has an impact on the performance of each individual model, but not on the overall ranking of each algorithm. Hence, we can state that the tuning of the hyper-parameters would not impact the conclusions of this dissertation.

A fifth limitation is that we do not apply transformations to our predictors. In Chapter 4 we apply a logarithmic transformation on the response variable but not on the predictors. Nevertheless variable transformations on continuous variables such as Box-Cox transformations have proven to be successful in social media optimization [214]. Other transformation on continuous (e.g., equal frequency discretization) and binary variables (e.g., weight of evidence) have also shown to significantly impact the performance of logistic regression models [51]. Future research could focus on implementing several variable transformation techniques and see whether they impact predictive performance.

A sixth limitation is the presence of selection effects. In Chapter 2 and 3 we gathered our data via a customized Facebook application for a European soccer team. This application was advertised several times on their Facebook page. To stimulate participation we offered a signed jersey. We also included a rules and regulations section and informed users that their data were extracted for academic purposes. To avoid privacy issues the extracted data were completely anonymous. Yet, we are aware that our data suffers from self-selection effects. First, since the application was advertised via the Facebook page of the European soccer, users who liked the Facebook page (or soccer in general) had a higher probability of being in our sample than users who did not. Users who did not like the Facebook page (or more in general soccer) could still see the application, however the sign-up rate would be rather low. Second, of those users who have seen the application in their News Feed, some will not be interested in the offered prize. Third, some

users may not be willing to share their data with Facebook or an academic institution. Finally, there is a social desirability bias that influences users to share things on Facebook. For example, users might share their attendance of an event or make their relationship 'Facebook official' because of the approval of their friends. In Chapter 4 we extracted the data from the official Facebook and Twitter page of a movie via the API. Also in this case there is a social desirability bias to comment or reply to a status update or a Tweet of the company. As a result we only model a subsample of the entire Facebook population and the results are not fully representative of the whole population. However, companies wanting to use social media data will always have the same limitations (e.g., they will have to advertise their application through the Facebook page or extract the data via the API, and these selection effects and social desirability biases will also occur). Therefore, this dissertation should be seen as a collection of valuable case studies on different levels of analysis and we are not claiming any generalization to the whole Facebook population. We hope that this dissertation will inspire other researchers in the field to replicate our results. If more researchers collect their own social media data sets and there are a sufficient number of case studies, a more conclusive answer can be offered by doing a meta-analysis [107].

A final limitation for Chapter 2 and 3 is the fact that some predictors have a restricted number of values. Facebook bounds the number of values that can be extracted per variable. For example, at the time of our data collection it was only allowed to extract the 25 most recent entries. To cope with this problem, we calculated the frequency of each variable within a specific time period, so that there is no entry in our data set that reaches this limit. For status updates, photo uploads, and likes created we computed the frequency of the last 7 days, for albums uploads and check-ins for the last 4 months, and for notes and video uploads for the last year.

Since the other limitations are specific to each study, we summarize these limitations in the following paragraphs.

5.3.2 Main limitations of each study

Chapter 2 is limited because we only include friends variables that are related to the focal event. A possible avenue for future research could be to include more friends variables and investigate which type of friends variables lead to the biggest increase in predictive performance following the example of Zhang et al. [228]. However, preliminary tests with extra friends variables did not reveal a significant increase in predictive performance.

A first limitation of **Chapter 3** is that in contrast to other studies about romantic partnerships [14] we do not include variables related to the whole social network of the user (i.e., topological features). The reason is that we were only

able to collect the ego networks of the users (i.e., only first degree friends). We also did not include these features since we are interested in interactions between ego and alter and how these difference in interaction influence their social tie. To state it differently, we decided to focus on the properties of the relationship and communication between ego and alter and not on the properties of the network [114]. Second, one of the most important indicators of romantic partnership is private communication or messages [40]. However, privacy regulations do not allow us to extract (the volume of) private communication on Facebook. We wanted to focus on observable interaction between ego and alter, since this information is also available for marketers. Hence, we believe that our disaggregated features capture the whole range of observable interaction between users on Facebook. Third, we do not study the traditional tie strength problem, namely estimating the actual tie strength rate or modeling whether or not someone is a strong or weak tie [100, 125]. Instead we measure whether or not ego is an alter's significant other and vice versa. An interesting avenue for future research would be to see whether our disaggregated variables have the same effect on tie strength. Closely related to this limitation is that we are actually modeling social ties. However, a lot of different social ties exist besides being a significant other (e.g., family relationships or colleagues). An interesting avenue for future research would be to study the impact of our disaggregated variables on different social ties and see whether our results hold across social ties.

Chapter 4 is limited because we do not include the genre or budget of the movie. The reason is that we wanted to estimate the average effect on box office sales using a diverse set of movies. An avenue for future research could be to redo our analysis for different types of movies (e.g., different genres and different budgets). Second, we decide to only model gross box office revenues as this is the most general measure of box office sales. Following the example of Ding et al. [68], it could be interesting to model opening weekend or opening month box office revenues. In line with Lash and Zhao [141] another interesting option would be to model movie success. Finally, we only use social media data to predict box office sales. However, features related to the characteristics of the movie (e.g., the cast and director) and the theaters (e.g., number of screens) have proven to have a significant impact on movie sales [141]. Moreover, other social media data could also have an impact on movie sales, such as YouTube, Yahoo! Movies and Google Trends. However, we decided to only focus on the two most important social media platforms in terms of users, namely Facebook and Twitter.

Bibliography

[1] Alpaydin, E., 1998. Combined 5 x 2 cv F test for comparing supervised classification learning algorithms. Neural Computation 11, 1885–1892.

[2] Altman, N., 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician 46 (3), 175–185.

[3] 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.

[4] Aral, S., Muchnik, L., Sundararajan, A., Dec. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proceedings of the National Academy of Sciences 106 (51), 21544–21549.

[5] Aral, S., Walker, D., Apr. 2014. Tie Strength, Embeddedness, and Social Influence: A Large-Scale Networked Experiment. Management Science 60 (6), 1352–1370.

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

[7] Arnaboldi, V., Conti, M., Passarella, A., Pezzoni, F., Sep. 2012. Analysis of Ego Network Structure in Online Social Networks. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). pp. 31–40.

[8] Arnaboldi, V., Conti, M., Passarella, A., Pezzoni, F., Apr. 2013. Ego networks in Twitter: An experimental analysis. In: 2013 Proceedings IEEE INFOCOM. pp. 3459–3464.

[9] Arnaboldi, V., Guazzini, A., Passarella, A., Jun. 2013. Egocentric online social networks: Analysis of key features and prediction of tie strength in Facebook. Computer Communications 36 (10–11), 1130–1144.

[10] Asur, S., Huberman, B. A., Aug. 2010. Predicting the Future with Social Media. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). Vol. 1. pp. 492–499.

[11] Baatarjav, E.-A., Amin, A., Dantu, R., Gupta, N., Jan. 2010. Are you my friend? [Twitter response estimator]. In: 2010 7th IEEE Consumer Communications and Networking Conference (CCNC). pp. 1–5.

[12] Backstrom, L., 2006. Group formation in large social networks: membership, growth, and evolution. In: In KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, pp. 44–54.

[13] Backstrom, L., 2013. News Feed FYI: A Window Into News Feed. URL `https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed`

[14] Backstrom, L., Kleinberg, J., 2014. Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing. ACM, New York, NY, USA, pp. 831–841.

[15] 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 (Supplement C), 13–23.

[16] Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., Dedene, G., Apr. 2002. Bayesian neural network learning for repeat purchase modelling in direct marketing. European Journal of Operational Research 138 (1), 191–211.

[17] Ballings, M., Van den Poel, D., Dec. 2012. Customer event history for churn prediction: How long is long enough? Expert Systems with Applications 39 (18), 13517–13522.

[18] Ballings, M., Van den Poel, D., Jun. 2013. Kernel Factory: An ensemble of kernel machines. Expert Systems with Applications 40 (8), 2904–2913.

[19] Ballings, M., Van den Poel, D., 2015. CRM in social media: Predicting increases in Facebook usage frequency. European Journal of Operational Research 244 (1), 248–260.

[20] Ballings, M., Van den Poel, D., Mar. 2015. R-package interpretR: Binary Classifier Interpretation Functions.

[21] Ballings, M., Van Den Poel, D., Sep. 2015. R-package kernelFactory: Kernel Factory: An Ensemble of Kernel Machines.

[22] Ballings, M., Van den Poel, D., May 2015. R-package rotationForest: Fit and Deploy Rotation Forest Models.

[23] Ballings, M., Van den Poel, D., Bogaert, M., Mar. 2016. Social media optimization: Identifying an optimal strategy for increasing network size on Facebook. Omega 59, Part A, 15–25.

[24] Ballings, M., Van den Poel, D., Hespeels, N., Gryp, R., Nov. 2015. Evaluating Multiple Classifiers for Stock Price Direction Prediction. Expert Systems with Applications 42 (20), 7046–7056.

[25] Baym, N. K., Ledbetter, A., Apr. 2009. Tunes That Bind? Information, Communication & Society 12 (3), 408–427.

[26] Benoit, D. F., Van den Poel, D., Oct. 2012. Improving customer retention in financial services using kinship network information. Expert Systems with Applications 39 (13), 11435–11442.

[27] Bentley, J. L., 1975. Multidimensional Binary Search Trees Used for Associative Searching. Commun. ACM 18 (9), 509–517.

[28] Berk, R. A., 2008. Statistical learning from a regression perspective. Springer.

[29] Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., 2013. R-package FNN: Fast Nearest Neighbor Search Algorithms and Applications. URL `https://cran.r-project.org/web/packages/FNN/index.html`

[30] Bhagwat, S., Goutam, A., 2013. Development of social networking sites and their role in business with special reference to Facebook. Journal of Business and Management 6 (5), 15–28.

[31] Bogaert, M., Ballings, M., Hosten, M., Van den Poel, D., 2017. Identifying Soccer Players on Facebook Through Predictive Analytics. Decision Analysis 14 (4), 274–297.

[32] 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.

[33] BoxOfficeMojo, 2017. Box Office Mojo.
URL `http://www.boxofficemojo.com/`

[34] Breiman, L., Aug. 1996. Bagging predictors. Machine Learning 24 (2), 123–140.

[35] Breiman, L., Oct. 2001. Random Forests. Machine Learning 45 (1), 5–32.

[36] Breiman, L., 2002. Manual On Setting Up, Using, And Understanding Random Forests V3.1. Statistics Department University of California Berkeley, CA, USA.

[37] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., 1984. Classification And Regression Trees. Wadsworth & Brookes/Cole Advanced Books & Software, USA.

[38] Burez, J., Van den Poel, D., 2007. CRM at a pay-TV company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. Expert Systems with Applications 32 (2), 277–288.

[39] Burez, J., Van den Poel, D., 2009. Handling class imbalance in customer churn prediction. Expert Systems with Applications 36 (3), 4626–4636.

[40] Burke, M., Kraut, R. E., 2014. Growing Closer on Facebook: Changes in Tie Strength Through Social Network Site Use. In: Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 4187–4196.

[41] 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 (4), 321–331.

[42] Chang, J., Sun, E., 2011. Location 3: How users share and respond to location-based data on social networking sites. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. pp. 74–80.

[43] 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.

[44] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321–357.

[45] Chen, J., Geyer, W., Dugan, C., Muller, M., Guy, I., 2009. Make New Friends, but Keep the Old: Recommending People on Social Networking Sites. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 201–210.

[46] 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 (5), 944–957.

[47] Choi, J.-H., Kang, D.-o., Jung, J., Bae, C., Oct. 2014. Investigating correlations between human social relationships and online communications. In: 2014 International Conference on Information and Communication Technology Convergence (ICTC). pp. 736–737.

[48] Coppens, S., Mannens, E., Pessemier, T. D., Geebelen, K., Dacquin, H., Deursen, D. V., Walle, R. V. d., Feb. 2011. Unifying and targeting cultural activities via events modelling and profiling. Multimedia Tools and Applications 57 (1), 199–236.

[49] Coussement, K., Benoit, D. F., Antioco, M., Nov. 2015. A Bayesian approach for incorporating expert opinions into decision support systems: A case study of online consumer-satisfaction detection. Decision Support Systems 79, 24–32.

[50] Coussement, K., De Bock, K. W., Sep. 2013. Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning. Journal of Business Research 66 (9), 1629–1636.

[51] Coussement, K., Lessmann, S., Verstraeten, G., Mar. 2017. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems 95, 27–36.

[52] Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S., 2008. Feedback effects between similarity and social influence in online communities. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 160–168.

[53] Cui, R., Gallino, S., Moreno, A., Zhang, D. J., 2017. The Operational Value of Social Media Information. Production and Operations Management.

[54] Culp, M., Johnson, K., Michailidis, a. G., Jun. 2012. ada: an R package for stochastic boosting.

[55] Dag, A., Oztekin, A., Yucel, A., Bulur, S., Megahed, F. M., 2017. Predicting heart transplantation outcomes through data analytics. Decision Support Systems 94 (Supplement C), 42–52.

[56] Daly, E. M., Geyer, W., 2011. Effective Event Discovery: Using Location and Social Information for Scoping Event Recommendations. In: Proceedings of the Fifth ACM Conference on Recommender Systems. ACM, New York, NY, USA, pp. 277–280.

[57] De Bock, K. W., Coussement, K., Van den Poel, D., Jun. 2010. Ensemble classification based on generalized additive models. Computational Statistics & Data Analysis 54 (6), 1535–1546.

[58] De Bock, K. W., Van den Poel, D., 2011. An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. Expert Systems with Applications 38 (10), 12293–12301.

[59] De Meo, P., Ferrara, E., Fiumara, G., Provetti, A., Oct. 2014. On Facebook, Most Ties Are Weak. Commun. ACM 57 (11), 78–84.

[60] 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 (5), 1–15.

[61] de Vries, L., Gensler, S., Leeflang, P. S. H., May 2012. Popularity of Brand Posts on Brand Fan Pages: An Investigation of the Effects of Social Media Marketing. Journal of Interactive Marketing 26 (2), 83–91.

[62] Dellarocas, C., Zhang, X. M., Awad, N. F., 2007. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. Journal of Interactive Marketing 21 (4), 23–45.

[63] DeLong, E. R., DeLong, D. M., Clarke-Pearson, D. L., Sep. 1988. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 44 (3), 837.

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

[65] Dietterich, T. G., Oct. 1998. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation 10 (7), 1895–1923.

[66] Dietterich, T. G., Jun. 2000. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. Springer Berlin Heidelberg, pp. 1–15.

[67] Díez-Pastor, J. F., Rodríguez, J. J., García-Osorio, C., Kuncheva, L. I., Sep. 2015. Random Balance: Ensembles of variable priors classifiers for imbalanced data. Knowledge-Based Systems 85, 96–111.

[68] 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.

[69] Dredge, S., 2014. Facebook squeezes 'overly promotional page posts' out of news feeds.
URL http://www.theguardian.com/technology/2014/nov/17/facebook-page-posts-news-feeds

[70] Dreiseitl, S., Ohno-Machado, L., Oct. 2002. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics 35 (5–6), 352–359.

[71] Du, J., Xu, H., Huang, X., 2014. Box office prediction based on microblog. Expert Systems with Applications 41 (4, Part 2), 1680–1689.

[72] 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.

[73] Duda, R. O., Hart, P. E., Stork, D. G., Nov. 2012. Pattern Classification. John Wiley & Sons.

[74] Dudoit, S., Fridlyand, J., Speed, T. P., 2002. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association 97 (457), 77–87.

[75] Dunbar, R. I. M., Arnaboldi, V., Conti, M., Passarella, A., Oct. 2015. The structure of online social networks mirrors those in the offline world. Social Networks 43, 39–47.

[76] Dunbar, R. I. M., Spoors, M., Sep. 1995. Social networks, support cliques, and kinship. Human Nature 6 (3), 273–290.

[77] Dunn, O. J., Mar. 1961. Multiple Comparisons among Means. Journal of the American Statistical Association 56 (293), 52–64.

[78] 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.

[79] Eliashberg, J., Hui, S. K., Zhang, Z. J., 2007. From Story Line to Box Office: A New Approach for Green-Lighting Movie Scripts. Management Science 53 (6), 881–893.

[80] 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 (4), 1143–1168.

[81] Facebook, 2014. News Feed FYI | Facebook Newsroom. URL https://newsroom.fb.com/news/category/news-feed-fyi/

[82] Facebook, 2015. Products | Facebook Newsroom. URL http://newsroom.fb.com/products/

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

[84] Facebook, 2017. Company Info | Facebook Newsroom. URL http://newsroom.fb.com/company-info/

[85] Facebook, 2017. Graph API - Documentation. URL https://developers.facebook.com/docs/graph-api/

[86] Facebook, 2018. Facebook-advertising. URL https://nl-nl.facebook.com/business/products/ads

[87] Fang, X., Hu, P. J.-H., Li, Z. L., Tsai, W., Jan. 2013. Predicting Adoption Probabilities in Social Networks. Information Systems Research 24 (1), 128–145.

[88] Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., Oct. 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research 15, 3133–3181.

[89] Festinger, L., 1954. A theory of social comparison processes. Human relations 7 (2), 117–140.

[90] Fournier, S., Rietveld, B., Dec. 2013. Get More Value Out of Social Media Brand-Chatter. Harvard Business Review.

[91] Freund, Y., Schapire, R. E., others, 1996. Experiments with a new boosting algorithm. In: ICML. Vol. 96. pp. 148–156.

[92] Friedman, J., Hastie, T., Simon, N., Tibshirani, R., Apr. 2015. R-package glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.

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

[94] Friedman, J. H., Feb. 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis 38 (4), 367–378.

[95] Friedman, J. H., Meulman, J. J., May 2003. Multiple additive regression trees with application in epidemiology. Statistics in Medicine 22 (9), 1365–1381.

[96] Friedman, M., 1940. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. The Annals of Mathematical Statistics 11 (1), 86–92.

[97] Gaikar, D. D., Marakarkandy, B., Dasgupta, C., 2015. Using Twitter data to predict the performance of Bollywood movies. Industrial Management & Data Systems 115 (9), 1604–1621.

[98] García, S., Fernández, A., Luengo, J., Herrera, F., May 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences 180 (10), 2044–2064.

[99] Gilbert, E., 2012. Predicting Tie Strength in a New Medium. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work. ACM, New York, NY, USA, pp. 1047–1056.

[100] Gilbert, E., Karahalios, K., 2009. Predicting Tie Strength with Social Media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 211–220.

[101] 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 (1), 88–107.

[102] Granovetter, M. S., 1973. The strength of weak ties. American journal of sociology, 1360–1380.

[103] Groh, G., 2007. Recommendations in Taste Related Domains: Collaborative Filtering vs. Social Filtering. In: In Proc ACM Group'07. pp. 127–136.

[104] Guàrdia-Sebaoun, É., 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. OAIR '13. Le Centre de Hautes Etudes Internationales d'Informatique Documentaire, Paris, France, France, pp. 201–208.

[105] Han, X., Wang, L., Crespi, N., Park, S., Cuevas, Á., Jan. 2015. Alike people, alike interests? Inferring interest similarity in online social networks. Decision Support Systems 69, 92–106.

[106] Hanley, J. A., McNeil, B. J., Apr. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143 (1), 29–36.

[107] Hanssens, D. M., Jan. 2018. The value of empirical generalizations in marketing. Journal of the Academy of Marketing Science 46 (1), 6–8.

[108] Hartmann, W. R., Manchanda, P., Nair, H., Bothner, M., Dodds, P., Godes, D., Hosanagar, K., Tucker, C., Dec. 2008. Modeling social interactions: Identification, empirical methods and policy implications. Marketing Letters 19 (3-4), 287–304.

[109] Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., Tibshirani, R., 2009. The elements of statistical learning. Springer.

[110] He, H., Garcia, E. A., Sep. 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21 (9), 1263–1284.

[111] 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 (3), 375–394.

[112] Hern, A., May 2018. Cambridge Analytica: how did it turn clicks into votes? The Guardian.
URL http://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-w

[113] Hernandez-Orallo, J., Flach, P., Ferri, C., Oct. 2012. A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss. Journal of Machine Learning Research 13, 2813–2869.

[114] Hill, R. A., Dunbar, R. I. M., Mar. 2003. Social network size in humans. Human Nature 14 (1), 53–72.

[115] Hong, L., Dan, O., Davison, B. D., 2011. Predicting Popular Messages in Twitter. In: Proceedings of the 20th International Conference Companion on World Wide Web. ACM, New York, NY, USA, pp. 57–58.

[116] Horvitz, E., Koch, P., Kadie, C. M., Jacobs, A., 2002. Coordinate: Probabilistic Forecasting of Presence and Availability. In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 224–233.

[117] Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., Schölkopf, B., 2009. Nonlinear causal discovery with additive noise models. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems 21. Curran Associates, Inc., pp. 689–696.

[118] Hu, T., Tripathi, A., 2016. The Performance Evaluation of Machine Learning Classifiers on Financial Microblogging Platforms. In: Internetworked World. Lecture Notes in Business Information Processing. Springer, Cham, pp. 74–83.

[119] Ingram, David, Mar. 2018. Factbox: Who is Cambridge Analytica and what did it do? Reuters.
URL https://www.reuters.com/article/us-facebook-cambridge-analytica-factbox/factbox-who-is-cambridge-analytica-and-what-did-it-do-idUSKB

[120] Itoga, H., Lin, G. T., others, 2013. Using Facebook for event promotion-implementing change. African Journal of Business Management 7 (28), 2788–2793.

[121] Jain, V., 2013. Prediction of movie success using sentiment analysis of tweets. The International Journal of Soft Computing and Software Engineering 3 (3), 308–313.

[122] James, G., Witten, D., Hastie, T., Tibshirani, R., Aug. 2013. An Introduction to Statistical Learning: with Applications in R, 1st Edition. Springer, New York.

[123] Janitza, S., Strobl, C., Boulesteix, A.-L., 2013. An AUC-based permutation variable importance measure for random forests. BMC Bioinformatics 14, 119.

[124] Jeners, N., Nicolaescu, P., Prinz, W., 2012. Analyzing Tie-Strength across Different Media. In: Herrero, P., Panetto, H., Meersman, R., Dillon, T. (Eds.), On the Move to Meaningful Internet Systems: OTM 2012 Workshops. Springer Berlin Heidelberg, pp. 554–563.

[125] Jones, J. J., Settle, J. E., Bond, R. M., Fariss, C. J., Marlow, C., Fowler, J. H., Jan. 2013. Inferring Tie Strength from Online Directed Behavior. PLoS ONE 8 (1), e52168.

[126] Kahanda, I., Neville, J., Mar. 2009. Using Transactional Information to Predict Link Strength in Online Social Networks. In: ICWSM. pp. 74–81.

[127] Kalampokis Evangelos, Tambouris Efthimios, Tarabanis Konstantinos, 2013. Understanding the predictive power of social media. Internet Research 23 (5), 544–559.

[128] Kayaalp, M., Ozyer, T., Ozyer, S. T., 2009. A Collaborative and Content Based Event Recommendation System Integrated With Data Collection Scrapers and Services at a Social Networking Site. Ieee, New York.

[129] Kemp, S., Apr. 2014. Global Social Media Users Pass 2 Billion. URL `http://wearesocial.net/blog/2014/08/global-social-media-users-pass-2-billion/`

[130] Kim, T., Hong, J., Kang, P., 2015. Box office forecasting using machine learning algorithms based on SNS data. International Journal of Forecasting 31 (2), 364–390.

[131] Klamma, R., Cuong, P. M., Cao, Y., Jan. 2009. You Never Walk Alone: Recommending Academic Events Based on Social Network Analysis. In: Zhou, J. (Ed.), Complex Sciences. Springer Berlin Heidelberg, pp. 657–670.

[132] Kossinets, G., Watts, D. J., Jun. 2006. Empirical Analysis of an Evolving Social Network. Science 311 (5757), 88–90.

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

[134] 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 (1), 7–25.

[135] Kwok, L., Yu, B., Feb. 2013. Spreading Social Media Messages on Facebook: An Analysis of Restaurant Business-to-Consumer Communications. Cornell Hospitality Quarterly 54 (1), 84–94.

[136] 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.

[137] Langley, P., 2000. Crafting Papers on Machine Learning. In: In Proceedings of the Seventeenth International Conference on Machine Learning (ICML). pp. 1207–1212.

[138] Langley, P., Iba, and, W., Thompson, K., 1992. An Analysis of Bayesian Classifiers. In: Proceedings of the Tenth National Conference on Artificial Intelligence. AAAI Press, San Jose, California, pp. 223–228.

[139] Langley, P., Sage, S., 1994. Induction of Selective Bayesian Classifiers. In: Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 399–406.

[140] Larivière, B., Van den Poel, D., Aug. 2005. Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications 29 (2), 472–484.

[141] Lash, M. T., Zhao, K., 2016. Early predictions of movie success: the who, what, and when of profitability. Journal of Management Information Systems 33 (3), 874–903.

[142] Lee, D. H., 2008. PITTCULT: Trust-based Cultural Event Recommender. Assoc Computing Machinery, New York.

[143] Lee, W., Xiong, L., Hu, C., Sep. 2012. The effect of Facebook users' arousal and valence on intention to go to the festival: Applying an extension of the technology acceptance model. International Journal of Hospitality Management 31 (3), 819–827.

[144] Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., Christakis, N., Oct. 2008. Tastes, ties, and time: A new social network dataset using Facebook.com. Social Networks 30 (4), 330–342.

[145] Li, L., Apr. 2018. Predicting online invitation responses with a competing risk model using privacy-friendly social event data. European Journal of Operational Research.

[146] Liao, H.-Y., Chen, K.-Y., Liu, D.-R., Jan. 2015. Virtual friend recommendations in virtual worlds. Decision Support Systems 69, 59–69.

[147] Liaw, A., Wiener, M., 2002. Classification and Regression by randomForest. R news 2 (3), 18–22.

[148] Lin, N., Dayton, P. W., Greenwald, P., Jan. 1978. Analyzing the Instrumental Use of Relations in the Context of Social Structure. Sociological Methods & Research 7 (2), 149–166.

[149] 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 (3), 1509–1528.

[150] Liu, X., Shen, H., Ma, F., Liang, W., Dec. 2014. Topical Influential User Analysis with Relationship Strength Estimation in Twitter. In: 2014 IEEE International Conference on Data Mining Workshop (ICDMW). pp. 1012–1019.

[151] Liu, Y., 2006. Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. Journal of Marketing 70 (3), 74–89.

[152] Liu, Y., Huang, X., An, A., Yu, X., 2007. ARSA: A Sentiment-aware Model for Predicting Sales Performance Using Blogs. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, New York, NY, USA, pp. 607–614.

[153] Lo, J.-P., 2010. The effectiveness of WOM by using Facebook as an implementation in movie industry. Ph.D. thesis, Califronia State Univeristy, Sacramento.

[154] Lovett, T., O'Neill, E., Irwin, J., Pollington, D., 2010. The Calendar As a Sensor: Analysis and Improvement Using Data Fusion with Social Networks and Location. In: Proceedings of the 12th ACM International Conference on Ubiquitous Computing. ACM, New York, NY, USA, pp. 3–12.

[155] Mallipeddi, R., Janakiraman, R., Kumar, S., Gupta, S., 2017. The Effects of Social Media Tone on Engagement: Evidence from Indian General Election 2014. SSRN Scholarly Paper ID 2980481, Social Science Research Network, Rochester, NY.

[156] Mangold, W. G., Faulds, D. J., Jul. 2009. Social media: The new hybrid element of the promotion mix. Business Horizons 52 (4), 357–365.

[157] Marsden, P. V., Campbell, K. E., Jan. 1984. Measuring Tie Strength. Social Forces 63 (2), 482–501.

[158] Marshall, Manson, Mar. 2014. Facebook Zero: Considering Life After the Demise of Organic Reach.
URL https://social.ogilvy.com/facebook-zero-considering-life-after-the-demise-of-organic-r

[159] Martens, D., Provost, F., Clark, J., Fortuny, E. J. d., Dec. 2016. Mining Massive Fine-Grained Behavior Data to Improve Predictive Analytics. Management Information Systems Quarterly 40 (4), 869–888.

[160] McPherson, M., Smith-Lovin, L., Cook, J. M., 2001. Birds of a feather: Homophily in social networks. Annual review of sociology, 415–444.

[161] Meire, M., Ballings, M., Van den Poel, D., 2017. The added value of social media data in B2b customer acquisition systems: A real-life experiment. Decision Support Systems 104 (Supplement C), 26–37.

[162] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Jul. 2015. R-package e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

[163] Minkov, E., Charrow, B., Ledlie, J., Teller, S., Jaakkola, T., 2010. Collaborative future event recommendation. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM Press, p. 819.

[164] Mishne, G., 2006. Predicting movie sales from blogger sentiment. In: In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW). pp. 155–158.

[165] Montgomery, D. B., Morrison, D. G., 1973. A Note on Adjusting R2. The Journal of Finance 28 (4), 1009–1013.

[166] Mooij, J., Janzing, D., 2008. Distinguishing Between Cause and Effect. In: Proceedings of the 2008th International Conference on Causality: Objectives and Assessment - Volume 6. COA'08. JMLR.org, Whistler, Canada, pp. 147–156.

[167] Moon, S., Bergey, P. K., Iacobucci, D., Jan. 2010. Dynamic Effects Among Movie Ratings, Movie Revenues, and Viewer Satisfaction. Journal of Marketing 74 (1), 108–121.

[168] Mynatt, E., Tullio, J., 2001. Inferring Calendar Event Attendance. In: Proceedings of the 6th International Conference on Intelligent User Interfaces. ACM, New York, NY, USA, pp. 121–128.

[169] Nemenyi, P., 1963. Distribution-free Multiple Comparisons. Princeton University.

[170] Ng, A. Y., Jordan, M. I., 2002. On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes. In: Dietterich, T. G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Information Processing Systems 14. MIT Press, pp. 841–848.

[171] Novet, J., Feb. 2014. Facebook's Valentine's Day gift to all of us: data about our relationships.
URL http://venturebeat.com/2014/02/15/facebooks-valentines-day-gift-to-all-of-us-data-about-our-re

[172] Ogata, H., Yano, Y., Furugori, N., Jin, Q., Jun. 2001. Computer Supported Social Networking For Augmenting Cooperation. Computer Supported Cooperative Work (CSCW) 10 (2), 189–209.

[173] 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 (1), 25–37.

[174] Oztekin, A., Al-Ebbini, L., Sevkli, Z., Delen, D., 2018. A decision analytic approach to predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology. European Journal of Operational Research 266 (2), 639–651.

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

[176] 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 (3), 697–710.

[177] Pan, R. K., Sinha, S., 2010. The statistical laws of popularity: Universal properties of the box office dynamics of motion pictures. ArXiv e-prints 1010, arXiv:1010.2634.

[178] Pappalardo, L., Rossetti, G., Pedreschi, D., Aug. 2012. 'How Well Do We Know Each Other?' Detecting Tie Strength in Multidimensional Social Networks. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 1040–1045.

[179] Paris, C. M., Lee, W., Seery, P., 2010. The Role of Social Media in Promoting Special Events: Acceptance of Facebook 'Events'. Springer-Verlag Wien, Vienna.

[180] Pennacchiotti, M., Popescu, A.-M., 2011. A Machine Learning Approach to Twitter User Classification. ICWSM 11, 281–288.

[181] Pessemier, T. D., Coppens, S., Geebelen, K., Vleugels, C., Bannier, S., Mannens, E., Vanhecke, K., Martens, L., May 2012. Collaborative recommendations with content-based filters for cultural activities via a scalable event distribution platform. Multimedia Tools and Applications 58 (1), 167–213.

[182] Peters, A., Hothorn, T., Ripley, B. D., Therneau, T., Atkinson, B., 2017. R-package ipred: Improved Predictors.

URL https://cran.r-project.org/web/packages/ipred/index.html

[183] Piatetsky, G., 2015. CRISP-DM, still the top methodology for analytics, data mining, or data science projects.
URL https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-html

[184] Prinzie, A., Van den Poel, D., Apr. 2008. Random Forests for multiclass classification: Random MultiNomial Logit. Expert Systems with Applications 34 (3), 1721–1732.

[185] Prinzie, A., Van den Poel, D. V. d., 2007. Random Multiclass Classification: Generalizing Random Forests to Random MNL and Random NB. In: Wagner, R., Revell, N., Pernul, G. (Eds.), Database and Expert Systems Applications. Springer Berlin Heidelberg, pp. 349–358.

[186] Recio-García, J. A., Quijano, L., Díaz-Agudo, B., Dec. 2013. Including social factors in an argumentative model for Group Decision Support Systems. Decision Support Systems 56, 48–55.

[187] 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 (1).

[188] Ridgeway, G., 2017. R-package gbm: Generalized Boosted Regression Models.
URL https://cran.r-project.org/web/packages/gbm/index.html

[189] Ripley, B., Venables, W., 2015. R-package nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models.
URL https://cran.r-project.org/web/packages/nnet/index.html

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

[191] Roberts, S. G. B., Dunbar, R. I. M., Pollet, T. V., Kuppens, T., May 2009. Exploring variation in active network size: Constraints and ego characteristics. Social Networks 31 (2), 138–146.

[192] Rodriguez, J., Kuncheva, L., Alonso, C., Oct. 2006. Rotation Forest: A New Classifier Ensemble Method. IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (10), 1619–1630.

[193] Rui, H., Liu, Y., Whinston, A., 2013. Whose and what chatter matters? The effect of tweets on movie sales. Decision Support Systems 55 (4), 863–870.

[194] Sandri, M., Zuccolotto, P., Jan. 2006. Variable Selection Using Random Forests. In: Zani, P. S., Cerioli, P. A., Riani, P. M., Vichi, P. M. (Eds.), Data Analysis, Classification and the Forward Search. Springer Berlin Heidelberg, pp. 263–270.

[195] Servia-Rodríguez, S., Díaz-Redondo, R. P., Fernández-Vilas, A., Blanco-Fernández, Y., Pazos-Arias, J. J., Apr. 2014. A tie strength based model to socially-enhance applications and its enabling implementation: mySocialSphere. Expert Systems with Applications 41 (5), 2582–2594.

[196] Sevim, C., Oztekin, A., Bali, O., Gumus, S., Guresen, E., Sep. 2014. Developing an early warning system to predict currency crises. European Journal of Operational Research 237 (3), 1095–1104.

[197] Shahbaznejad, H., Dolan, R., Tripathi, A., 2017. The Power of Facebook and Instagram Fans: An Exploration of Fan Comments and Their Effect on Social Media Content Strategy. SSRN Scholarly Paper ID 3054984, Social Science Research Network, Rochester, NY.

[198] Sharda, R., Delen, D., 2006. Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications 30 (2), 243–254.

[199] Sheng, D., Sun, T., Wang, S., Wang, Z., Zhang, M., 2013. Measuring Strength of Ties in Social Network. In: Ishikawa, Y., Li, J., Wang, W., Zhang, R., Zhang, W. (Eds.), Web Technologies and Applications. Springer Berlin Heidelberg, pp. 292–300.

[200] Singla, P., Richardson, M., 2008. Yes, There is a Correlation: - from Social Networks to Personal Behavior on the Web. In: Proceedings of the 17th International Conference on World Wide Web. ACM, New York, NY, USA, pp. 655–664.

[201] Spackman, K. A., 1991. Maximum likelihood training of connectionist models: comparison with least squares back-propagation and logistic regression. Proceedings of the Annual Symposium on Computer Application in Medical Care, 285–289.

[202] Spence, M., 1973. Job Market Signaling. The Quarterly Journal of Economics 87 (3), 355–374.

[203] Suh, B., Hong, L., Pirolli, P., Chi, E. H., Aug. 2010. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In: 2010 IEEE Second International Conference on Social Computing (SocialCom). pp. 177–184.

[204] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M., 2011. Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics 37 (2), 267–307.

[205] Therneau, T., Atkinson, B., Ripley, B., 2017. R-package rpart: Recursive Partitioning and Regression Trees.
URL `https://cran.r-project.org/web/packages/rpart/index.html`

[206] Thorleuchter, D., Van den Poel, D., Dec. 2012. Predicting e-commerce company success by mining the text of its publicly-accessible website. Expert Systems with Applications 39 (17), 13026–13034.

[207] Topuz, K., Uner, H., Oztekin, A., Yildirim, M. B., Apr. 2018. Predicting pediatric clinic no-shows: a decision analytic framework using elastic net and Bayesian belief network. Annals of Operations Research 263 (1-2), 479–499.

[208] Trattner, C., Steurer, M., Jul. 2015. Detecting partnership in location-based and online social networks. Social Network Analysis and Mining 5 (1), 1–15.

[209] Tullio, J., Mynatt, E. D., 2007. Use and Implications of a Shared, Forecasting Calendar. In: Baranauskas, C., Palanque, P., Abascal, J., Barbosa, S. D. J. (Eds.), Human-Computer Interaction – INTERACT 2007. Springer Berlin Heidelberg, pp. 269–282.

[210] Twitter, 2017. Bedrijf | About.
URL `https://about.twitter.com/nl/company`

[211] Twitter, 2017. Twitter Developer Platform — Twitter Developers.
URL `https://developer.twitter.com/en.html`

[212] Venkatesh, K., Ravi, V., Prinzie, A., Poel, D. V. d., Jan. 2014. Cash demand forecasting in ATMs by clustering and neural networks. European Journal of Operational Research 232 (2), 383–392.

[213] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B., Apr. 2012. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research 218 (1), 211–229.

[214] Wang, X., Baesens, B., Zhu, Z., 2018. On the Optimal Marketing Aggressiveness Level of C2c Sellers in Social Media: Evidence from China. Omega Forthcoming.

[215] Wei, J.-T., Lin, S.-Y., Wu, H.-H., 2010. A review of the application of RFM model. African Journal of Business Management 4 (19), 4199.

[216] Wellman, B., 2007. The network is personal: Introduction to a special issue of Social Networks. Social Networks 29 (3), 349–356.

[217] Wiese, J., Min, J.-K., Hong, J. I., Zimmerman, J., 2015. "You Never Call, You Never Write": Call and SMS Logs Do Not Always Indicate Tie Strength. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM, New York, NY, USA, pp. 765–774.

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

[219] Wolpert, D. H., Oct. 1996. The Lack of A Priori Distinctions Between Learning Algorithms. Neural Computation 8 (7), 1341–1390.

[220] 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. WOSN '12. ACM, New York, NY, USA, pp. 61–66.

[221] Xiang, R., Neville, J., Rogati, M., 2010. Modeling Relationship Strength in Online Social Networks. In: Proceedings of the 19th International Conference on World Wide Web. ACM, New York, NY, USA, pp. 981–990.

[222] Xie, K., Lee, Y.-J., 2015. Social Media and Brand Purchase: Quantifying the Effects of Exposures to Earned and Owned Social Media Activities in a Two-Stage Decision Making Model. Journal of Management Information Systems 32 (2), 204–238.

[223] Xie, Y., Chen, Z., Zhang, K., Jin, C., Cheng, Y., Agrawal, A., Choudhary, A., 2013. Elver: Recommending Facebook Pages in Cold Start Situation Without Content Features. Ieee, New York.

[224] Xu, K., Zou, K., Huang, Y., Yu, X., Zhang, X., Jan. 2016. Mining community and inferring friendship in mobile social networks. Neurocomputing 174, Part B, 605–616.

[225] Xu, Y., Guo, X., Hao, J., Ma, J., Lau, R. Y. K., Xu, W., Dec. 2012. Combining social network and semantic concept analysis for personalized academic researcher recommendation. Decision Support Systems 54 (1), 564–573.

[226] Zanda, A., Eibe, S., Menasalvas, E., Jul. 2012. SOMAR: A SOcial Mobile Activity Recommender. Expert Systems with Applications 39 (9), 8423–8429.

[227] Zhang, H., Dantu, R., May 2010. Predicting social ties in mobile phone networks. In: 2010 IEEE International Conference on Intelligence and Security Informatics (ISI). pp. 25–30.

[228] Zhang, Y., Wu, H., Sorathia, V., Prasanna, V. K., Jul. 2013. Event Recommendation in Social Networks with Linked Data Enablement. In: Proceedings of 15th International Conference on Enterprise Information Systems. pp. 371–379.

[229] Zhao, J., Wu, J., Liu, G., Tao, D., Xu, K., Liu, C., Oct. 2014. Being rational or aggressive? A revisit to Dunbar's number in online social networks. Neurocomputing 142, 343–353.

[230] Zhao, X., Yuan, J., Li, G., Chen, X., Li, Z., Oct. 2012. Relationship strength estimation for online social networks with the study on Facebook. Neurocomputing 95, 89–97.

[231] Zou, K. H., Tuncali, K., Silverman, S. G., 2003. Correlation and simple linear regression. Radiology 227 (3), 617–622.