Predicting Donation Behavior: Acquisition Modeling in the Nonprofit Sector Using Facebook Data

Lisa Schetgen[a], Matthias Bogaert[a], Dirk Van den Poel[a]

[a]Ghent University, Department of Marketing, Tweekerkenstraat 2, 9000 Ghent, Belgium

Lisa.Schetgen@UGent.Be, Matthias.Bogaert@UGent.Be (Corresponding author),

Dirk.VandenPoel@UGent.Be

Abstract

The purpose of this study is to demonstrate the value of Facebook data in predicting first-time donation behavior. More specifically, we provide evidence that Facebook data can be used as a valuable data source for nonprofit organizations in acquiring new donors. To do so, we evaluate three different dimensionality reduction techniques (i.e., singular value decomposition, non-negative matrix factorization, and latent Dirichlet allocation) over seven classification techniques (i.e., logistic regression, k-nearest neighbors, bagged trees, random forest, adaboost, extreme gradient boosting, and artificial neural networks) using five times twofold cross-validation. Next, we assess what type of Facebook data and which predictors are most important. The results indicate that we can predict first-time donation behavior based on Facebook data with high predictive performance. Our benchmark indicates that the combination of singular value decomposition and logistic regression outperforms all other analytical methodologies with an area under the receiver operating characteristic of 0.72 and a top decile lift of 3.33. The results show that Facebook pages and categories of Facebook pages are the most important data types. The most important predictors are dimensions related to age, education, residence, materialism, responsible consumption, and interest in nonprofits. The presented acquisition models can be used by nonprofit organizations to implement a one-to-one targeted marketing campaign towards Facebook fans. To the best of our knowledge, our study is the first to determine the predictive value of Facebook data for nonprofits in a real-life acquisition context.

Keywords: Facebook, CRM, Customer Acquisition, Predictive Analytics, Social Media

1. Introduction

Nonprofit organizations (NPOs) are characterized by a strong devotion to their end users, namely those in need. To successfully fulfill their social mission, acquiring potential donors and retaining existing donors is crucial. Hence, customer relationship management (CRM), which focuses on the relationship with donors, can be extremely valuable to them. For example, Verhaert and Van den Poel [60] demonstrate how nonprofits can use transactional data in order to predict which individuals are likely to donate again. It goes without saying that successful CRM initiatives can lead to increased loyalty of donors, significant cost reductions, and improved fundraising results [40]. In recent years, the relationship between nonprofit organizations and donors has been dramatically transformed by the emergence of social media. Due to limited knowledge and resources, most nonprofit organizations use these platforms for communication purposes [26]. They are rarely used as a source of valuable information or as a tool for decision-making. Nonprofit organizations realize that social media have an impact on CRM, yet they remain unsure on how to incorporate them into their CRM strategies and activities [46].

Traditionally, it is very hard for nonprofit organizations to find new potential donors. Customer acquisition is inherently difficult, as organizations need information about donors whom they have no existing relationship with [58]. Therefore, in order to identify potential donors, NPOs often have no other choice but to acquire expensive external databases. However, the emergence of social media could empower organizations to identify potential customers. By liking a NPO's Facebook page, users publicly display a certain interest in the organization and can, therefore, be seen as new potential donors. Hence, by extracting the list of individuals who have liked their social media page, nonprofits can pinpoint a substantial base of prospects [58]. Furthermore, social media contains a huge amount of behavioral data on these potential donors. In sum, instead of using external data sources, an organization's own social media data could be

used to construct an acquisition model, which identifies users that are most likely to donate. Subsequently, these users can directly be targeted with tailored campaigns. Despite the opportunities of social media for donor acquisition, no study evaluates how social media data can be used for acquisition modeling in the nonprofit sector. Nevertheless, this could be worthwhile for a multitude of reasons. First, even though academics and practitioners generally focus on customer retention, acquisition of new donors is vital to any nonprofit organization. Considering the limited customer lifetime of donors [52], there will always be a certain number of donors leaving the organization and the necessity to acquire new ones. Second, the (added) value of social media data in prospecting, and CRM in general, has already been proven in several applications [8,49]. Because of the shrinking donor pool and the fierce competition amongst organizations, it is important for nonprofits to make their acquisition models as accurate as possible [35,53]. Finally, social media data can be a valuable source for mining information with regards to personal characteristics and behavior [38,58]. The latter are relevant to nonprofit organizations as they can be employed to improve targeted marketing and fundraising efforts. Given the lack of studies focusing on predicting donation behavior with social media data, several questions remain unanswered: (1) "Is it possible to predict donation behavior using solely social media data, and if so, what data analytical methodologies are required?", (2) "What type of social media data matters the most?", and (3) "Which variables are most important?".

To fill this gap in literature, this paper investigates whether social media data (i.e., Facebook data) can be used to predict donation behavior. More specifically, we construct an acquisition model aimed at predicting which potential donors (i.e., Facebook fans of the organization) are most likely to become actual donors. To do so, we worked together with a well-known European nonprofit organization and gathered publicly available Facebook data of their potential donors (i.e., liked Facebook pages, categories of liked Facebook pages, joined Facebook groups, frequency variables, and gender). To effectively predict donation behavior, we come up with a decision support system to build acquisition models based on social media data. The goal

of this decision support system is to provide a data collection and analytical framework such that NPOs and/or researchers can easily replicate and implement our approach. Our data collection framework describes the steps to gather the prospect list from the NPO's Facebook page and collect the behavioral data from these prospects on Facebook. Our analytical framework consists of a specific two-stage process. First, we evaluate different dimensionality reduction techniques (i.e., singular value decomposition (SVD), non-negative matrix factorization (NMF), and latent Dirichlet allocation (LDA)) to cope with the great number of Facebook pages and groups. To increase the robustness of our results, we benchmark these different dimensionality reduction techniques over seven prediction algorithms (i.e., logistic regression (LR), k-nearest neighbors (KNN), bagged trees (BT), random forest (RF), adaboost (AB), extreme gradient boosting (XGB), and artificial neural networks (NN)). We contribute to existing literature by determining the optimal combination of data reduction and modeling techniques when using social media data. This optimal combination then serves as the input of the second stage of our process, in which we determine the most valuable Facebook features. To determine the (added) value of different variable types, we construct several models on different subsets of variables. For instance, we compare the performance of the most complete model to a model that excludes variables related to Facebook page categories. Finally, we also assess variable importances to uncover the driving forces of predictive performance.

The remainder of this paper is organized as follows. First, we provide an overview of existing literature. Second, we explain our methodology in more detail. Next, we present an overview of our results, followed by a conclusion and discussion of their practical implications. Finally, we formulate recommendations for future research.

2. Literature overview

Since an elaborate overview of social CRM literature would be too extensive, we focus our literature review on CRM studies in the nonprofit sector. In other words, we review literature

that explores donation behavior from a CRM point-of-view. Taking into account the relevance of social media to our study, we also include studies about donation behavior using social media.

Table 1: Overview of CRM literature concerning the nonprofit sector, as well as studies regarding donation behavior using social media

Study	**Research type**		**CRM domain**		**Data type**	
	Diagnostic	Predictive	Retention	Acquisition	Traditional	Social media
Sargeant [53]	x		x		x	
Bennett [7]	x		x	x	x	
Hsu et al. [35]	x		x	x	x	
Germain et al. [32]	x		x		x	
Ferguson [27]	x		x	x	x	
Schlumpf et al. [54]	x		x		x	
Garner & Garner [31]	x		x		x	
Enjolras et al. [24]	x				x	x
Chell & Mortimer [15]	x		x		x	x
Brown & Taylor [12]	x				x	x
Courtois & Verdegem [17]	x				x	x
Warren et al. [61]	x					x
Farrow & Yuan [26]	x				x	x
Godin et al. [33]		x	x		x	
Althoff & Leskovec [2]		x	x		x	
Lee & Chang [42]		x	x	x	x	
Verhaert & Van den Poel [60]		x	x	x	x	
Our study		**x**		**x**		**x**

Table 1 categorizes the existing literature according to three dimensions: (1) research type, (2) CRM domain, and (3) type of data. First, two research types can be distinguished. Diagnostic research typically investigates underlying reasons and potential determinants of the behavior of interest, in our case donation behavior. Specifically, these studies examine how different factors (e.g., personal values, satisfaction, and motivating factors) impact donor return [31,54], recency and frequency of donation behavior [27], and the choice of charity [7]. The derived results are inherently historical as behavior is described and explained based on information from the past. On the other hand, the goal of predictive research is to determine how current information can be used to predict future behavior. For example, Althoff and Leskovec [2] use logistic regression models to predict whether donors on online crowdfunding platforms will donate again. Second, literature can be divided according to the application in the CRM domain. Studies that focus on customer retention investigate potential reasons for churn or repeated donation behavior, and

explain how their results can be incorporated into retention efforts (e.g., [33]). Studies focusing on customer acquisition highlight how their analyses and results can be used to identify potential donors and work out specific marketing strategies (e.g., [60]). Finally, research on donation behavior can be categorized according to the type of data. Traditional data consists of past donation behavior (e.g., [35]), socio-demographic data (e.g., [32]), personality traits (e.g. [27]), and intentions (e.g., [15]). Social media data includes self-reported information about social media usage, as well as information directly retrieved from social media websites. Most studies that use social media data use the former. For example, Brown and Taylor [12] employ self-reported information to determine whether there is a relationship between presence on social media and donation behavior. Enjolras et al. [24], as well as Farrow and Yuan [26], take it one step further by investigating the underlying reasons of the positive impact of using social media on donating time and money. Chell and Mortimer [15] take a CRM point-of-view and explore the value of granting online recognition on social media to motivate individuals to donate again. With regards to the type of data, Courtois and Verdegem [18] are the only ones who use directly gathered social media data in addition to self-reported data. In their study, they retrieve information related to Facebook pages and groups (e.g., the specific goal and the number of page likes) in order to investigate the determinants of connective action on Facebook.

From Table 1, it is clear that no study has built a predictive model on the basis of solely social media data for acquisition purposes in the nonprofit sector. Our contribution to existing literature is twofold. First, none of the predictive studies in Table 1 are performed in a real-life acquisition context. In the study by Lee and Chang [42], the dependent variable (i.e., whether someone is likely to donate) is derived from self-reported information. Furthermore, no distinction is made between first-time donations and redonations. In contrast to Lee and Chang [42], Verhaert and Van den Poel [60] model *actual* donation behavior as their dependent variable. However, they focus on existing donors only, which implies that they model redonation instead of first-time donation. They extrapolate their results to an acquisition context by suggesting that

empathy could be a valuable predictor for new donors. Hence, we are the first to investigate the prediction of donation behavior in a real-life acquisition context. Second, despite the fact that research in the nonprofit sector has evolved from self-reported to crawled social media data, all of the studies inherently perform diagnostic research. For example, Brown and Taylor [12] and Enjolras et al. [24] perform a regression analysis to determine the effects of specific individual attributes on charitable behavior. Farrow and Yuan [26] use structural equation modeling to test causal hypotheses and study the relationship between using a NPO's Facebook group and actual behavior. Another downside is that there is only one study that is situated in the CRM domain. Chell and Mortimer [15] use Pearson correlation and regression analysis techniques to test whether online recognition has an impact on someone's intention to donate again. Their results can assists NPOs to develop online retention strategies that leverage self-interest over altruism but they do not assess whether their findings can be used to predict future donations. Hence, there are no predictive studies on acquisition modeling using social media data in the nonprofit sector. This is a missed opportunity since Facebook has over 2 billion active monthly users [25] and social media data is available for (almost) everyone. Hence, this would allow organizations to access an enormous number of potential donors. Moreover, when exclusively using social media data, nonprofit organizations do not need to have a pre-existing relationship with potential donors to collect information and work out acquisition initiatives. In other words, social media allows decision makers to perform real-life acquisition experiments.

To fill this gap in literature, we create a decision support system for constructing an acquisition model on the basis of Facebook data. The aim of this model is to predict whether a fan of a nonprofit organization's Facebook page is likely to become an actual donor. Nowadays, nonprofit organizations are forced to use their own databases in order to build predictive models. Consequently, they can only focus on their own customer base. Leveraging the power of their Facebook pages (and social media data in general), allows them to identify new potential donors (i.e., users who like their Facebook page), target those Facebook fans who are most likely to

become a donor, and gain insights into their characteristics. To deliver a decision support system that allows NPOs to predict acquisition of their Facebook fans, we present a framework for the collection of relevant Facebook data (Facebook pages, categories of Facebook pages, Facebook groups, and socio-demographics) and employ a specific data analytical process. In the first step of this process, we benchmark three dimensionality reduction techniques over seven classification algorithms to find the optimal combination of both methods. Second, we use this optimal combination to evaluate different subsets of variables. This allows us to establish the (added) value of the different types of Facebook features. Furthermore, we compute variable importances to discover which features are most important in predicting donation behavior.

There are several reasons why we believe Facebook can be valuable for the development of donor acquisition models. First, Facebook can represent an important acquisition platform as it enables new opportunities of gaining and sharing information [24], which can strengthen the relationship with prospective donors [26] and ultimately the decision to donate [35]. Furthermore, Courtois and Verdegem [17] state that Facebook allows individuals to find and interact with people supporting the same good cause and can consequently engender feelings of positive group acceptance and identification. Farrow and Yuan [26] later confirmed these findings for Facebook groups. Hence, considering Facebook makes it possible to identify a novel pool of interested individuals (i.e., Facebook fans) and has the potential of solidifying the relationship with these individuals, it would be a missed opportunity not to investigate this new source of potential donors. Second, we believe that Facebook data is valuable for the prediction of first time donation behavior as previous research on social media data for prospecting confirms the potential value of Facebook data [8,49]. On the basis of Facebook data, Bogaert et al. [8] are able to accurately predict which individuals practice a certain hobby (i.e., soccer). Sports brands can then use this information to identify new potential customers and decide which ones to target. Whereas, Bogaert et al. [8] focus on a B2C setting, Meire et al. [49] investigate the added value of Facebook data for customer acquisition in a B2B setting. They find that models incorporating

Facebook data are considerably better at predicting good prospects. Hence, as Facebook data has proven to be a valuable predictor in several acquisition situations, we believe it could be equally valuable for acquisition modeling in the nonprofit sector. Third, previous work on predicting user behavior with social media data has shown that individual Facebook features can be used to accurately predict personal characteristics such as age, gender, relationship status, and personality traits [3,37,38]. Considering that these characteristics are also significantly related to the act of giving [6,12,14,52], we believe that Facebook data can similarly be used for the prediction of donation behavior. Specifically, there are two types of Facebook features that we suspect to be particularly valuable for this task. Kosinski et al. [38] demonstrate the value of Facebook pages in predicting a wide range of private attributes that are related to charitable giving (e.g., age, personality traits, political viewpoints). The value of Facebook likes results from the fact that an ever-increasing proportion of human activities takes place online. Hence, Facebook likes represent digital records of human behavior and are used by users to express their positive association with online content [38]. Besides Facebook pages, we believe that the categories of these liked pages could be valuable for the prediction of donation behavior as well. Zhang and Pennacchiotti [62] find that categories of liked Facebook pages hold enough information to predict users' purchase behavior. This could be explained by the fact that Facebook users like specific categories and as such express personal interests [62]. Based on the findings in existing literature, there are strong indications that Facebook pages and categories of Facebook pages will be amongst the top variables when predicting donation behavior.

To summarize, our study does not only allow us to determine whether it is possible to predict donation behavior with Facebook data, but also enables us to gain insights into the most important data types and predictors. Moreover, on the basis of previous literature, we find strong indications that donation behavior (i.e., becoming a new donor) can be accurately predicted with individual Facebook data. To the best of our knowledge this is the first study to investigate the feasibility of acquisition modeling on the basis of Facebook data in the nonprofit sector.

3. Methodology

3.1. Data

To gather our Facebook data, a data collection framework similar to van Dam and van de Velden [58] is used. The first step consists of identifying users who liked the NPO's Facebook page. In what follows, we will refer to these users as fans. A list of the page fans can be viewed by going to https://www.facebook.com/12345/settings/?tab=people_and_other_pages, where 12345 should be replaced by the NPO's Facebook page ID. It is important to note that this URL is exclusively accessible to administrators of the concerned Facebook page. In February 2018, a list of 8646 unique fans was retrieved by scraping the aforementioned Facebook page using the RSelenium R-package [34]. Fans' Facebook names, along with their Facebook ID and the date on which they started to like the NPO's Facebook page, were extracted from the webpage's source code using regular expressions. The second step of the data collection framework proposed by van Dam and van de Velden [58] involves the retrieval of relevant Facebook data for each of the identified fans. In the context of this research, the relevant data consists of liked Facebook pages, joined Facebook groups, and socio-demographics (i.e., gender and age). As this data cannot be collected using Facebook's Graph API, an alternative manner was pursued. Individuals' pages and groups were collected by visiting and scraping the following webpages: https://www.facebook.com/browse/fanned_pages/?id=12345&title=Pagina+die+ik+leuk+vind&starttime=0&endtime=6789 and https://www.facebook.com/search/12345/groups&starttime=0&endtime=6789. In these URL's, 12345 should be replaced by a fan's Facebook ID and 6789 should represent the date on which this particular fan started to like the NPO's Facebook page. The first URL leads to a webpage displaying the Facebook pages that an individual has liked up until the moment of liking the NPO's Facebook page. Similarly, the second URL gives access to a webpage that consists of a list of Facebook groups joined by a particular individual. In essence, for each of the 8646 unique

Facebook ID's, the aforementioned webpages were accessed and scraped between February and May 2018. By applying regular expressions to the source code of the respective webpages, we were able to extract a list of liked Facebook pages, containing the names of the pages as well as the corresponding categories, and a list of joined Facebook groups, consisting of the groups' names. For the collection of gender and age, a similar approach was employed. Between May and June 2020, we scraped the following page for every fan in our sample and extracted information on gender and birthdate using regular expressions: https://www.facebook.com/12345/about?section=contact-info, where 12345 should be replaced by the fan's Facebook ID. It is important to mention that only publicly available information is collected[1]. In other words, the collected data exclusively concerns users who allow public access to their Facebook information.

Since it is imperative that most information is available for all observations in our sample, we delete Facebook fans for whom no data is available about Facebook pages and/or groups. As over 95% of Facebook users in our sample allow public access to information related to pages and groups, this reduces our sample to 8246 observations. Furthermore, we exclusively consider Facebook pages and groups that have been liked or joined prior to liking the NPO's Facebook page. This is crucial for the correctness of our predictive model. When developing a predictive model, the predictors should be computed based on data from the independent period [41]. Whereas the independent period has no predefined start date, it ends at the time of liking the NPO's Facebook page, as shown in Figure 1. As such, all the Facebook pages (groups) that an individual has liked (joined) up until the moment of liking the NPO's Facebook page are included in the computation of the predictors. The dependent variable (i.e., donor or not) should be computed based on data from the dependent period, meaning the period following after the act of liking the Facebook page (see Figure 1). Hence, the dependent period starts from the moment a Facebook fan likes the NPO's Facebook page and ends on the 8th of April 2018 (i.e., the date we

[1] Facebook's privacy regulations state that public information can be viewed, opened, shared, and downloaded by third parties. See https://www.facebook.com/full_data_use_policy for more information.

received a copy of the internal donor database of the NPO). As the end of the independent period and the start of the dependent period are different for every Facebook fan in our sample, the length of the dependent period ranges from 39 to 715 days.

Besides Facebook data we also possess transactional data from the NPO, which will be used to identify actual donors amongst the Facebook fans. The data is merged via names, as this is the only piece of information common to both datasets [50]. To sum up, for each individual in our sample we have the following data (if available): (i) name, (ii) date of becoming a donor (if applicable), (iii) date of liking the NPO's Facebook page, (iv), list of the names of liked Facebook pages, (v) list of the categories of liked Facebook pages, (vi) list of the names of joined Facebook groups, and (vii) socio-demographics (gender and birthdate). To guarantee correct matching, names occurring more than once are deleted from our sample. The resulting sample consists of 8167 Facebook fans who have liked 667,740 unique Facebook pages, corresponding to 1738 unique page categories, and joined 213,633 unique Facebook groups. Of those 8167 Facebook fans, we matched 477 individuals to the NPO's internal database.

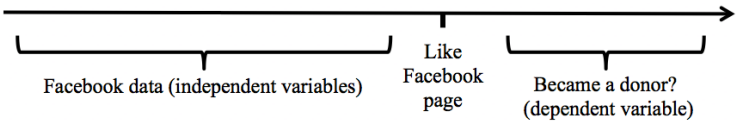

Figure 1: Time window

3.2. Variables

Dependent variable

Our dependent variable is a binary variable that indicates whether a Facebook fan has become a donor (1) or not (0). To make sure that we do not violate the time window of our predictive model, the dependent variable is computed based on information from the period *after* liking the NPO's Facebook page (see Figure 1). Furthermore, an acquisition model is aimed at *prospective* donors. This means that the individuals in our sample cannot be donors at the end of the independent period. Hence, our data sample should consist of the Facebook fans that are

targetable. These are either Facebook fans that are not donors at all, or Facebook fans that became a donor *after* liking the NPO's Facebook page.

In practice, the selection of relevant observations and the computation of the dependent variable are based on the date of becoming a donor and the date of liking the NPO's Facebook page. If the first variable is missing, this means that the Facebook user could not be matched to the internal data and thus is not a donor. Hence, the dependent variable is set to 0. In case the date of first donation takes place between the date of liking the Facebook page and the 8th of April 2018, the user is identified as a donor and the dependent variable is set to 1. Finally, user that were already a donor before the liking the NPO's Facebook page, are excluded from our sample.

Our final data sample consists of 7795 observations. From these, only 96 became a donor in the dependent period (i.e., after liking the NPO's Facebook page). In other words, merely 1,23% (96 out of 7795) of the individuals in our sample are donors, whereas 98,77% (7699 out of 7795) are not. This class imbalance is inherent to the issue at hand. Namely, on social media, the number of fans and followers of nonprofit organizations is ever increasing but only very few of them actually donate. Hence, identifying these few potential donors is of great value and is exactly what our predictive model is aimed at. To cope with the class imbalance problem, we follow the recommendation of Bogaert et al. [10] to use random oversampling as their study finds that it is superior to other data sampling methods on social media data (i.e., undersampling and SMOTE). Following their recommendations, we performed oversampling on the training set until a 50/50 distribution was reached.

Independent variables

For the construction of the independent variables, we start by creating matrices corresponding to the different Facebook features: (1) user-page matrix, (2) user-category matrix, and (3) user-group matrix. The rows of the matrices correspond to the Facebook fans, whereas the columns correspond to the liked Facebook pages, the categories, and the joined Facebook groups.

For example, element x_{ij} of the user-page matrix is set to 1 if user *i* likes Facebook page *j*. The same rationale can be applied to the other matrices. 68,59% of the Facebook pages (78,33% of the Facebook groups) are liked (joined) by only one individual in our sample. Based on the recommendation of Kosinki et al. [38,39], we decide to keep only those pages, categories, and groups that have been liked by at least 20 users in our sample. This reduces the number of unique Facebook pages, page categories, and Facebook groups to respectively 11,408, 1098, and 1402.

Nevertheless, the dimensionalities and sparsity levels of our matrices remain high. Therefore, we apply several dimensionality reduction (DR) techniques to these matrices. The main reason for using DR in predictive modeling is to increase predictive performance while reducing sampling variance [16]. Furthermore, DR could uncover latent information present in our Facebook data. We decide to perform DR on each of the matrices separately, as this allows us to distinguish between the (predictive value of the) different types of Facebook features in the second stage of our research. The different dimensionality reduction techniques are presented in Section 3.3.

Furthermore, we also include the frequency variables into our predictive model: the number of liked Facebook pages and the number of joined Facebook groups. For each Facebook fan, these are calculated as the sum of the corresponding row in the user-page matrix and the user-group matrix respectively. Finally, we consider key socio-demographical variables, namely gender and age. The variable gender was successfully collected for 6571 out of the 7795 Facebook users in our final data sample. The remaining 1224 values are imputed based on the fans' first names using the Genderize.io API[2]. With regards to the variable age, information on birthdate was only available for 7.99% of the Facebook fans in our sample. Another option could be to deduce age by linking users' Facebook ZIP codes with census data. Unfortunately, similarly to age, this information is available for very few Facebook users. Given the large number of missing values, we discarded the variable age. In sum, the independent variables for our

[2] https://genderize.io/

predictive model consist of *k* dimensions related to Facebook pages, page categories, and Facebook groups, which result from the application of a DR technique on the respective matrices (see Section 3.3), as well as the number of Facebook pages, the number of Facebook groups, and the user's gender.

Models

One of the goals of this study is to assess the value of the different types of Facebook features. To do so, we create different models summarized in Table 2. Models 1 (M1), 2 (M2), and 3 (M3) are, respectively, built on Facebook pages, Facebook page categories, and Facebook groups only. This is done to investigate whether any of these types of Facebook features on their own are sufficient to create an accurate prediction model. Model 5 (M5) consists of the frequency variables, the variables related to Facebook pages, as well as those related to Facebook page categories. In comparison with Model 8 (i.e., the most complete model), the variables related to Facebook groups have been left out in M5. By comparing the performance measures of Model 5 and Model 8, we are able to judge the added value of considering Facebook features related to groups. The same rationale can be applied to Model 4 (M4), Model 6 (M6), and Model 7 (M7). Note that the variable gender is added as a control variable to all models.

From Table 2, it is clear that we do not consider all possible combinations of Facebook features. This is because we are specifically interested in the (added) predictive value of the different types of Facebook data.

Table 2: Description of the models

	Description	M1	M2	M3	M4	M5	M6	M7	M8
Pages	100 dimensions related to Facebook pages, resulting from the application of a dimensionality reduction technique on the user-page matrix	x			x	x	x		x
Categories	100 dimensions related to categories of Facebook pages, resulting from the application of a dimensionality reduction technique on the user-category matrix		x		x	x		x	x

Groups							
	100 dimensions related to Facebook groups, resulting from the application of a dimensionality reduction technique on the user-group matrix	x	x		x	x	x
Frequency							
	#Liked pages' (total number of Facebook pages) and '#Groups' (total number of Facebook groups)			x	x	x	x

3.3. Dimensionality reduction

In this section we discuss the different dimensionality reduction (DR) techniques that we apply to our input data (i.e., user-page, user-category, and user-group). The following three DR techniques are used: singular value decomposition (SVD), non-negative matrix factorization (NMF), and latent Dirichlet allocation (LDA). SVD is the most popular dimensionality reduction technique in existing literature [38]. Its popularity can be attributed to its simplicity and computational speed [39]. Furthermore, Clark and Provost [16] compared the impact of several DR techniques on the performance of predictive models and found SVD to be the best technique. Just as SVD, NMF is a matrix factorization-based DR technique. However, in contrast to SVD, NMF requires the resulting components to be non-negative. Consequently, the number of components is often sparse and easy to interpret [16,51]. Finally, LDA is traditionally used as a clustering and topic modeling technique but can be equally useful for dimensionality reduction purposes. It is easy to interpret but can be computationally expensive and requires the original data to be non-negative [39]. A description of the different DR techniques is given in Table 3.

Choosing the optimal number of dimensions (k) is crucial when applying dimensionality reduction techniques. If k is too small, the concepts related to the resulting dimensions will be too broad. If k is too large, this may lead to similar concepts [51]. Several authors highlight the fact that there is no single correct way of identifying the optimal number of dimensions [16,39,51]. Kosinski et al. [39] note that this choice should depend on the application. For example, when using DR to build predictive models, it could be useful to choose a larger number of dimensions, as these will preserve more information from the original data. However, when k is set too high, the benefits of DR could be lost, leading to a decrease in accuracy [39].

Similarly to Praet et al. [51], we decide on the optimal value of *k* by iterating over several values and inspecting the resulting predictive performances. Specifically, we constructed a model for every combination of algorithm, DR technique and level of *k* (50, 100, and 200) (see Section 3.5). The optimal number of dimensions is chosen according to the value of *k* that results in the best performing model, measured by the AUC.

To assess whether using DR techniques such as SVD, NMF, and LDA, improves the predictive performance, it is imperative to compare these DR techniques with a dataset that has not undergone any feature engineering [16]. Hence, we also create a baseline model for which we do not apply any DR techniques to our input data. In this model the prediction variables are equal to the columns in the user-page, user-category, and user-group matrices. The columns represent binary variables that indicate whether a Facebook fan has liked a certain page, category or group (see Section 3.2). To ensure that we only keep the Facebook pages, categories and groups that are meaningful to the classification task (i.e., donor or not), we follow the advice of van Dam and van de Velden [58] to only maintain the top 100 Facebook pages, categories and groups in our baseline model. The top pages and page categories (top groups) are determined based on the number of likes (group members) in our sample.

Table 3: Description of dimensionality reduction techniques

DR technique	Description
Singular Value Decomposition	SVD represents the user-page matrix (A) as the product of three matrices ($A = U\Sigma V^T$): user-to-concept matrix (U), page-to-concept matrix (V), and a square matrix containing the singular values on the diagonal (Σ). The former two respectively represent how much a given user and Facebook page correspond to a given concept. The singular values in the latter are sorted in decreasing order of explained variance in the original data. By retaining the k first singular values, the *k* most important dimensions are identified [39].
Non-negative Matrix Factorization	NMF factorizes a given matrix V into two matrices W and H, under the constraint that all three matrices are positive: V≈WH. As the goal is to approximate V, W and H are computed such that the Frobenius norm of the difference V-WH is minimized. The columns in H, which contain weights, indicate how to reconstruct an approximation of the original data vectors as a combination of the basis elements in W [43].
Latent Dirichlet Allocation	LDA is based on the idea that Facebook users like a certain distribution of latent topics and that these topics are described by distributions of Facebook pages. The application of LDA results in the computation of two matrices ϕ and θ. ϕ represents the importance of Facebook pages per topics. θ indicates the importance of topics for the Facebook users. The probability p_{ij} that user i likes page j can be computed as [55]: $p_{ij} = \sum_{t=1}^{T} \phi_{jt}\, \theta_{ti}$

3.4. Prediction algorithms

Since we are interested in classifying Facebook fans into donors and non-donors, we employ several prediction algorithms. We chose algorithms that have been proven to yield superior performance in analytical CRM and social media (e.g., [10,49]) as this ensures the comparability of our results. The algorithms can be divided into two broad categories: single classifiers and ensemble techniques. The single classifiers are logistic regression (LR), k-nearest neighbors (KNN), and an artificial neural network (NN). LR is one of the most popular single classifiers and is often used as a benchmark model in analytical CRM [18,21]. It estimates the conditional probability $p(y \mid x)$ by maximizing the likelihood function [40]. One downside of LR is that it is prone to overfitting when introduced with a lot of variables [8]. Therefore we applied the LASSO (i.e., least absolute shrinkage and selection operator) approach [57]. Whereas LR is a parametric algorithm (i.e., it simplifies the learning function to a known form and learns the related coefficients from the training data), KNN is nonparametric. KNN makes no assumptions about the form of the mapping function other than that data points that are close are likely to have a similar outcome. It is equally a popular nonparametric method in social media applications [10]. Finally, NN is a semi-parametric method, as it requires to select a specific functional form a priori. The parameters of the model are estimated using nonlinear optimization [45]. NN has often been employed in social media analytics [9] and has achieved superior performance in large-scale benchmark studies [45] due to its ability to uncover complex nonlinear relationships between the predictors and the response variable. For our study, we train a NN with only one hidden layer and a logistic activation function using backpropagation since it has been proven that NNs with one hidden layer serve as universal approximators [4].

The ensemble techniques in this study are bagged trees (BT), random forest (RF), adaboost (AB), and extreme gradient boosting (XGB). These classifiers are often found to be top performers in both CRM [44] and social media studies [5]. Ensemble algorithms aggregate a set

of individual classifiers by combining their individual predictions [23]. Even though the considered algorithms are all tree-based ensembles, they differ in how they generate and aggregate the different decision trees. BT generates multiple classifiers by manipulating the training examples (i.e., taking bootstrap sample of the training data and aggregating the bootstrap samples). It is identified as one of the top performing ensembles in customer churn [44]. RF adds an extra layer of randomness to BT by selecting only a random subset of the predictors at each tree split [11]. RF was found to be the best performing algorithm in a number of social media applications, such as B2B acquisition [49]. Whereas BT and RF are parallel ensemble methods (i.e., the individual trees are generated in parallel), AB and XGB generate the individual learners sequentially. AB iteratively gives more weight to misclassified observations [28], whereas XGB minimizes the gradient of the loss function in each iteration [29]. Both AB [8,9,10] and XGB [19] have proven to be successful in social media applications. Table 4 summarizes the settings of the hyperparameters of the different prediction algorithms.

Table 4: Hyperparameter settings of the prediction algorithms

Prediction algorithm	Hyperparameter(s)	Settings
Logistic Regression (LR)	Regularization parameter (lambda)*	Fold-based: 0 < lambda < 0.25
K-Nearest Neighbors (KNN)	Number of nearest neigbors (k)*	10, 100, 200, …, 1000
Bagged Trees (BT)	Number of trees	500
Random Forest (RF)	Number of trees Number of predictor variables in the random subset at each node of the tree	500 sqrt(number of variables)
AdaBoost (AB)	Number of iterations Number of terminal nodes in the base classifiers	500 8
Extreme Gradient Boosting (XGB)	Maximum depth of a tree (max_depth)* The learning rate (eta)* The minimum loss reduction required to create a partition on a leaf (gamma)*	2, 3, 5, 7, 10, 100 0.025, 0.05, 0.1, 0.2, 0.3 0, 0.1, 0.2, 0.3, 0.4, 1.0, 1.5, 2.0
Neural Network (NN)	Number of hidden nodes (size)* Regularization parameter (decay)*	2, 3, …, 20 10^(-4, -3.5, …, 0)

* Hyperparameter was tuned.

3.5. Experimental set-up

To evaluate predictive performance we use two distinct measures that are widely accepted as appropriate evaluation metrics for the performance of classification algorithms, namely AUC (Area Under ROC Curve) and top decile lift (TDL) [48,49].

The advantage of the AUC, in comparison with other performance measures, is that it does not depend on a specific threshold [13]. The receiver operating characteristic (ROC) curve graphically represents the relation between the true positive (TP) rate and the false positive (FP) rate [10]. The AUC can be interpreted as the probability that a random positive observation gets a higher score than a random negative observation [13]. An AUC of 0.5 indicates that the predictive model does not perform any better than a random classifier. On the other hand, a value equal to 1 stipulates that the predictive model is perfect [9].

The top decile lift (TDL) measures how much better a predictive model is in identifying positives, in comparison to simply selecting instances at random. It is defined as the ratio of the percentage of positives (i.e., donors) in the top 10% of potential donors that have been attributed the highest scores and the overall percentage of positives [49].

The AUC and TDL measure two different aspects of model performance. Whereas the AUC measures the performance over the entire range of predictions, TDL focuses on the predictions with the highest likelihood to become a donor. Inspecting both of them, allows for a more nuanced evaluation of the predictive performance of the models [48].

To guarantee the robustness of our results, we perform five times twofold cross-validation (5x2fcv) [22]. The fist step is to randomly split the data into two distinct non overlapping samples of equal size (i.e., 50/50 distribution). Subsequently, the first sample is used to compute the new dimensions (i.e., resulting from the application of a specific DR technique for a pre-specified level of k) and to train the model (i.e., training set), whereas the second sample is used to evaluate its performance (i.e., test set), and vice versa. In case the algorithm requires hyperparameter tuning, the training set is again equally split into a training and a validation set, and a grid search is performed to select the optimal parameter(s). Subsequently, the model is re-trained on the full

training set with the optimal parameter settings. By repeating this process 5 times, we obtain 10 performance values [5]. As mentioned in Section 3.3, 5x2fcv is repeated for every combination of algorithm, DR technique and level of *k* (50, 100, and 200) separately. In other words, 700 models are compared in total: 7 algorithms and 3 DR techniques cross-validated 10 times and repeated for 3 levels of *k*, as well as 7 binary baseline models equally cross-validated 10 times. Finally, the reported results are the best combination of algorithm and DR, by taking the median AUC and TDL over the 10 cross-validation runs and selecting the optimal level of *k.*

To determine whether the performances of the different models are significantly different from each other, we make use of the combined Alpayadin F-test [1] with Holm corrections for family-wise error [30]. Existing literature covers a wide range of statistical tests for comparison of prediction models. However, they generally focus on comparing classifiers across different data sets [20]. As the aim of this study is to compare several models based on the same data set, these tests are not appropriate and we use the combined Alpayadin F-test instead.

4. Results

4.1. Dimensionality reduction and prediction algorithms

Table 5 provides an overview of the cross-validated model performance measures (i.e., AUC and TDL) for all possible combinations of DR techniques and prediction algorithms. The optimal level of *k* is reported between brackets[3] and varies across the different combinations of algorithms and DR techniques. The results indicate that acquisition of NPOs' Facebook fans can be accurately predicted on the basis of individual Facebook features: the AUC ranges from 0.55 to 0.72, depending on the employed techniques. The best performing models per algorithm are marked in bold in Table 5. We note that these do not differ significantly in terms of AUC. The top performer (underlined and in bold) is the combination of SVD and LR and results in an AUC of

[3] The optimal level of *k* is only included for the AUC in Table 5. The same levels of *k* were used for the TDL and therefore not reported.

0.72 and a TDL of 3.33. This means that the probability that a random donor gets a higher score than a random non-donor is equal to 0.72. In other words, the added value of our top-performing model over a random model is 0.22. Focusing on the 10% highest predictions, our model identifies 3.33 times more donors than a random model. The fact that the combination of LR and SVD is found to be the top performer could be due to the specificities of SVD. In the context of this study, SVD decomposes the user-page matrix into a user-to-concept matrix (displaying how much a given user corresponds to a given concept), a matrix containing the singular values on the diagonal (representing the strength of each concept), and a page-to-concept matrix (displaying to which extend a given Facebook page corresponds to a given concept). The singular values are positive and sorted in decreasing order of the amount of variance they account for in the original data. SVD operates as a DR technique by selecting the *k* first singular values (i.e., the dimensions that capture most of the variation in the original data) [39]. By excluding the low-variance components, SVD overcomes the multicollinearity problem [47]. In other words, SVD creates new dimensions that are linear combinations of the original features and that are orthogonal and uncorrelated by definition [47]. The absence of multicollinearity, which is one of the conditions of LR, is as such fulfilled. Moreover, applying SVD will also make sure that there is an approximate linear relationship between the predictors [47]. Hence, SVD will tackle one of the major downsides of regression techniques, namely their sensitivity to multicollinearity and nonlinearities in the data. Perhaps this is even more the case in the context of this research as liking a specific Facebook page is most likely strongly correlated with liking another page and the relationships are inherently nonlinear [5]. By applying SVD and taking away these correlations and nonlinearities, LR is able to estimate the true function.

Furthermore, based on the results in Table 5, we can state that, in comparison to the baseline model, the use of a DR technique generally leads to an increase in overall predictive performance (AUC). Across several prediction algorithms, SVD appears to be the best performing DR technique, which is in line with the findings of Clark and Provost [16]. In terms of TDL, the

application of a DR technique only leads to an important increase in performance in case of LR and XGB. Nevertheless, our results demonstrate the importance of dimensionality reduction as the top performer for both AUC and TDL can only be identified by evaluating all possible combinations of DR and prediction algorithms. To sum up, the top performer would have never been found if we had only considered binary variables (i.e., variables that have not undergone any feature engineering). Furthermore, additional analysis[4] revealed that applying DR techniques is necessary to make our models more stable and robust to overfitting. The binary models woefully overfit the data, which explains their suboptimal performance. Next, the optimal combination is used for the construction of different models and the computation of variable importances in the second part of this study.

Table 5: Cross-validated median AUC and TDL for all combinations of DR techniques and prediction algorithms

	AUC				TDL			
	Binary	SVD	NMF	LDA	Binary	SVD	NMF	LDA
LR	0.6658	**0.7239 *(k=100)***	0.6960 *(k=100)*	0.6989 *(k=100)*	2.6265	**3.3305**	2.9500	2.7250
KNN	0.6326	0.6078 *(k=100)*	**0.6408 *(k=100)***	0.6263 *(k=50)*	**2.0770**	1.5125	1.9980	1.7215
BT	0.5898	0.6155 *(k=100)*	0.6114 *(k=50)*	**0.6532 *(k=100)***	1.7130	1.6020	1.5550	**1.7700**
RF	0.6663	**0.6820 *(k=100)***	0.6735 *(k=50)*	0.6757 *(k=100)*	**2.4185**	1.8185	2.0020	2.1515
AB	0.6387	**0.6639 *(k=100)***	0.5496 *(k=50)*	0.5985 *(k=50)*	**2.3645**	1.8595	1.1920	1.5935
XGB	0.5978	**0.6630 *(k=100)***	0.5833 *(k=50)*	0.6463 *(k=50)*	1.7685	1.8300	1.2970	**2.0145**
NN	0.5854	0.6793 *(k=200)*	0.6624 *(k=200)*	**0.7009 *(k=50)***	**2.2985**	2.1365	1.9790	2.2080

Considering only Facebook data were included, as well as the difficulty of acquisition modeling, our model performance is very satisfying. Specifically, the performance of our best model (i.e., the combination of SVD and LR) is highly competitive compared to existing acquisition literature. Meire et al. [49], who modeled acquisition in a B2B setting, obtained AUCs ranging from 0.54 to 0.61. Similarly, D'Haen et al. [21] and Thorleuchter et al. [56] reached maximum AUCs of 0.62 and 0.61 respectively. Also in terms of TDL, our best model performs

[4] The results are available upon request.

well in comparison to previous acquisition studies. For example, Thorleuchter et al. [56] report TDL values ranging from 1.35 to 1.65.

As we consider Facebook data only, our results can be compared to existing literature in social media analytics. Specifically, we examine studies that are concerned with predicting personal characteristics based on social media data. In a study by Kosinski et al. [38], the AUCs vary between 0.6 and 0.95 depending on the characteristic that is being predicted. The best performing models correspond to the prediction of personal attributes such as gender or origin. On the other hand, the prediction of increasingly tricky characteristics, such as relationship status or use of drugs, consistently generate lower AUCs [38]. Considering the fact that donation behavior is a complex personal attribute, our results are in line with previous research and can be perceived as good.

4.2. Models

To assess the value of the distinct types of Facebook data, we constructed several models on different subsets of features (see Section 3.2). First, we evaluate the *individual* predictive value of each subset. Figure 2(a) shows the AUCs of the models build on Facebook pages, Facebook page categories, and Facebook groups only (M1, M2, and M3 respectively), as well as the AUC of the most complete model (M8). In terms of AUC, M1 and M2 appear to be comparable to M8. To determine whether this is statistically true, we inspect the results of the combined Alpayadin F-test with Holm corrections for family-wise error in Table 6. The lower diagonal of Table 6 presents the adjusted p-values for the differences in AUC, whereas the upper diagonal displays the p-values for the differences in TDL. The significant differences are marked in italic. The F-test confirms that, in terms of AUC, M1 and M2 do not perform significantly worse than the most complete model (M8). Moreover, we can state that Facebook pages alone (M1) are sufficient to create an accurate acquisition model. The same conclusion can be made for the categories of Facebook pages (M2). Besides inspecting the overall performance of these models, we also

evaluate how well they perform in terms of TDL. The resulting TDLs are presented in Figure 2(b). We find that, just as in terms of AUC, M1 and M2 perform as well as M8 (see Table 6). M3, which performs significantly worse than M8 at the 1% significance level in terms of AUC, only performs worse than M8 at the 10% significance level in terms of TDL (see Table 6). When comparing M1, M2, and M3, we find that M1 an M2 perform significantly better than M3 in terms of AUC, but not in terms of TDL. Even though the difference with M3 is not statistically significant, Figure 2(b) shows that M1 and M2 are also superior in terms of TDL. These results highlight the individual predictive power of Facebook pages and page categories for the prediction of donation behavior.

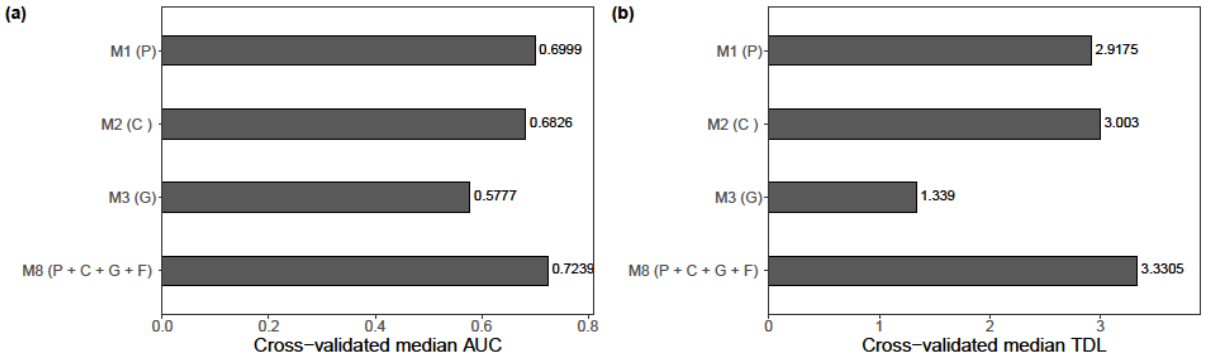

Figure 2: Cross-validated median (a) AUC and (b) TDL of M1, M2, M3, and M8

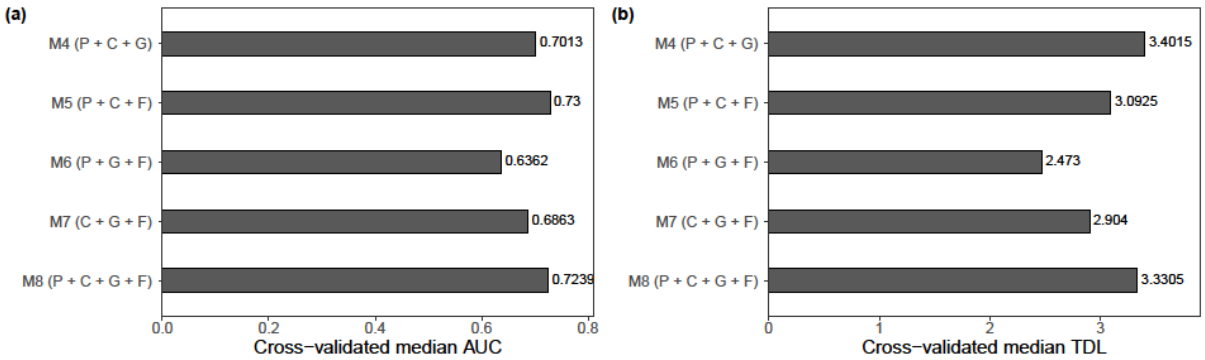

Figure 3: Cross-validated median (a) AUC and (b) TDL of M4, M5, M6, M7, and M8

Second, we evaluate the *added* value of the different types of Facebook features. By creating a model that excludes one type of features, and comparing its performance to that of the most complete model (M8), the added value of those features can be determined. Figure 3(a) displays the AUCs of the resulting models in comparison to the AUC of M8. We find that M6

(i.e., the model leaving out the Facebook page categories) has the lowest AUC, followed by M7 (i.e., the model leaving out the Facebook pages). In other words, categories of Facebook pages have an added value of 0.09 (= 0.7239 - 0.6362). Similarly, including features related to Facebook pages leads to an increase in AUC of 0.04 (= 0.7239 – 0.6863). The added value of Facebook groups and frequency variables appears to be smaller. It is important to note that M6 is the only model that performs significantly worse than M8 in terms of AUC (see Table 6), ratifying the added value of Facebook page categories. With regards to the TDL, even though the differences between the models are not statistically significant (see Table 6), Figure 3(b) shows that similar conclusions as for the AUC can be made.

Moreover, Table 6 shows that, with an AUC of 0.58, M3 (i.e., model including only Facebook groups) is by far the worst performing model in terms of AUC, as it is significantly different from all other models, except M6. For both AUC and TDL, M3 performs significantly worse than M4 at the 1% significance level. Hence, adding features related to Facebook pages and page categories leads to a significant increase in TDL from 1.34 to 3.40 (see Figures 2(b) and 3(b)). These findings suggest that, in comparison to Facebook groups, Facebook pages and their categories are especially valuable for the prediction of donation behavior of Facebook fans.

Table 6: Alpayadin F-test with Holm corrections

Adj. p		TDL							
		M1	M2	M3	M4	M5	M6	M7	M8
AUC	M1		0.7641	0.5710	0.7641	0.7641	0.7641	0.7641	0.7641
	M2	0.6832		0.3873	0.7641	0.7641	0.4459	0.7641	0.7641
	M3	*0.0138***	*0.0105***		*0.0096****	0.1670	0.4349	0.4353	*0.0844**
	M4	0.6832	0.6832	*0.0031****		0.7641	0.5710	0.7641	0.7641
	M5	0.6832	0.6832	*0.0001****	0.6832		0.4873	0.7641	0.7641
	M6	0.5025	0.6832	0.4763	0.6443	0.3448		0.7641	0.2676
	M7	0.6832	0.6283	*0.0099****	0.6832	0.6832	0.6832		0.7641
	M8	0.6832	0.6832	*0.0007****	0.6832	0.6832	*0.0973**	0.6832	

* $p < 0.1$,** $p < 0.05$, *** $p < 0.01$

The importance of Facebook pages and categories of Facebook pages is in line with existing literature regarding social media analytics. Kosinski et al. [38] have previously proven the predictive value of Facebook pages for a wide range of personal characteristics. With regards

to the categories of liked Facebook pages, Zhang and Pennacchiotti [62] found that these features solely are enough to accurately predict users' online purchase behavior. They state that this can be explained by the fact that Facebook users like specific categories in line with their personal interests.

4.3. Predictors

The final goal of this research is to determine which specific features are most important in the prediction of donation behavior. To do so, we examine variable importances in our most complete model (i.e., M8). For a regularized logistic regression, the importance of each variable can be measured by its corresponding coefficient (i.e., β coefficient). In contrast to other approaches, LASSO regression allows to determine which features exhibit the strongest effects [57], as coefficients that do not add any value to the performance of the model are shrunken towards zero. Hence, the smaller the coefficient, the smaller the importance of the corresponding variable. Figure 4 shows a scree plot of the 25 most important variables of M8. Specifically, this plot depicts the absolute value of the β coefficients in descending order on the y-axis, against their rank on the x-axis. The variable with the highest absolute coefficient receives rank 1, the one with the second highest absolute coefficient receives rank 2 and so on. The variables in our model are the SVD dimensions related to the user-page, user-category, and user-group matrices. The shape of the points in the variable importance plot represents the underlying type of Facebook data (i.e., pages are rhombuses, categories squares, groups triangles, frequency circles, and gender a star). Hence, Figure 4 allows us to confirm what type of Facebook data is most important in the prediction of donation behavior.

Figure 4 shows that the top predictor (i.e., the variable with rank 1) is related to categories of Facebook pages. Furthermore, we find that among the 25 most important variables 11 are related to Facebook pages, 10 to page categories, and only 4 to Facebook groups. These results confirm the importance of variables related to Facebook pages and page categories in the

prediction of donation behavior. Frequency variables do not appear in the top 25 and we can, therefore, conclude that these are less important. Similarly, gender is not one of the 25 most important variables.

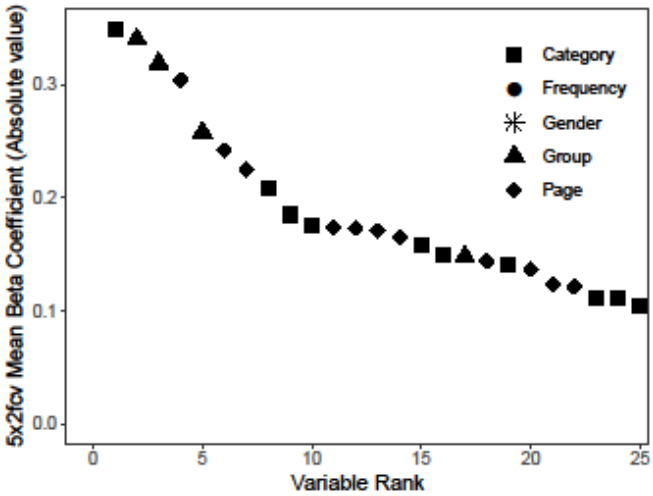

Figure 4: Variable importance plot of M8

Finally, we take a closer look at the underlying pages, categories and groups of the top ten most important variables. Recall that most of our variables are represented by the dimensions of SVD. Table 7 then displays the underlying Facebook pages/categories/groups that make up the top ten SVD dimensions, measured by the absolute value of the β coefficients for M8. By inspecting the former, we can gain insights into what distinguishes donors from non-donors and what aspects are most important for the prediction of donation behavior of Facebook fans. Most pages and groups in Table 7 have been translated from Dutch or French. If applicable, the original language is indicated, as this may hold valuable information for the interpretation of the dimensions. Furthermore, when deemed necessary for interpretation, extra information is given between brackets in italic.

Table 7 shows that the most important variable is strongly related to categories describing aspects of nightlife (e.g., Bar, Dance- and nightclub, and Pub) as well as education (e.g., Education, and Higher education and university). In other words, this dimension is most likely associated with students and young adults. These findings are in line with existing literature as it is generally accepted that the decision, as well as the amount to donate, are related to

characteristics such as age and education (e.g., [12,52]). The second most important variable in our model is a dimension related to second hand and sales groups and could, therefore, describe users that want to save money and/or consume responsibly. The third most important dimension clearly describes people from Liège, a city in the French-speaking part of Belgium. The fourth most important variable in the prediction of donation behavior is strongly related to Facebook pages of left-wing and humoristic media channels. The sixth most important dimension consists of Facebook pages of several well-known brands (e.g., CLUSE, Loavies, Coca-cola). Hence, this variable could be a measure for materialism and describe users that appreciate brands. Previous research by Bennett [7] shows that, in contrast to what one would expect, materialism positively influences the generosity of donations. Finally, the seventh most important variable of M8 is related to NPOs (e.g. Doctors Without Borders, Amnesty International, and UNICEF).

In summary, Facebook fans that become donors differ from non-donors in terms of their life stage, which part of Belgium they come from, the media channels they like, but also in terms of materialism, responsible consumption and interest in (other) NPOs. Our findings are in line with previous research on donation behavior, where demographics and socio-economic profiles [12,52], but also personal values and personality traits [7], are found to have an impact on the decision to donate.

Table 7: Top ten most important variables (i.e., SVD dimensions) of M8, along with the most related Facebook pages/categories/groups

1. Category	**2. Group**	**3. Group**
Bar	Namur, Belgium	Police Control Belgium**
Dance- and nightclub	2nd hand clothes NIVELLES**	Info radars, police controls Liège**
Pub	Selling, buying services in Brussels**	You are a real Liégois…**
Publisher	2nd hand and new items for sale, Antwerp*	Recommended restaurants in Liège**
Public service	Bart's way*	Liégeoises Liégeois**
Café	Halle Selling, Free or Trading group*	Radars Liège**
Education	For sale in the region of Mons**	Most beautiful pictures of Liège**
Higher education and university	For sale in Mons and its environs**	Bob from Liège**
Business service	All cars for sale in Belgium**	Gastronomy from Liège**
Jewelry/watches	AJITEBADLO.COM	Police, Accidents, Road Conditions**

4. Page	**5. Group**
Studio Brussel *(radio station)*	VIRTUAL FLEA MARKET LIEGE**
Stromae	Zoo of Antwerp*
Decovry.com - Be the first to	For sale Brussels**

discover	
De Ideale Wereld *(tv show)*	Looking for apartment(-sharing) Brussels**
Doctors Without Border (Belgium)*	JOBS BELGIUM
Humo *(magazine)*	Info radars and controls-Charleroi**
Nutella	Green bio zero waste**
De Morgen *(newspaper)*	Bouddhism
UNICEF Belgium	Second hand books Brussels**
UNILAD	2ND HAND CLOTHES CHARLEROI**

6. Page	**7. Page**	**8. Category**
Sneaker District	Doctors Without Borders (Belgium)**	Touring company
Decovry.com - Be the first to discover	Amnesty International Belgium**	Business service
Guts & Gusto	UNICEF Belgium	Café
CLUSE	Amnesty International	Magazine
Loavies	GuiHome vous détend *(humorist)*	Hairdresser
Colourful Rebel	Doctors Without Borders**	Publisher
Coca-Cola	UNICEF	Performance Art
TravelBird	Barack Obama	Higher education and university
LADbible	Doctors Without Border (Belgium)*	Medical company
EF Belgium - Study abroad	Doctors of the World Belgium**	Telecommunication company

9. Category	**10. Category**
Organization	Event planner
Public organization	Hairdresser
Society- and culturewebsite	Society- and culturewebsite
Cosmetics store	Education
Health- and wellnesswebsite	Youth organization
Magazine	Library
Public service	Non-governmental organization (NGO)
Other	Dance- and nightclub
Movie-/tv-studio	Sports
Entrepreneur	Fastfoodrestaurant

* Translated from Dutch, ** Translated from French

5. Conclusion and practical implications

This research shows how NPOs can harness the power of social media in CRM. Specifically, our aim is to demonstrate how nonprofit organizations can use social media (data) for acquisition modeling. First, we build a predictive model to determine whether it is *possible* to accurately predict donation behavior using Facebook data. Given the considerable number of Facebook pages, categories and groups, we evaluate different combinations of DR techniques (i.e., SVD, NMF, and LDA) and prediction algorithms (LR, KNN, BT, RF, AB, XGB, NN) to determine which combination of techniques is best suited for this task. Second, we build several

models based on different subsets of variables and computed variable importances to gain insights into the (most important) predictors of donation behavior.

The results indicate that acquisition of Facebook fans can be predicted on the basis of Facebook data with high predictive accuracy. With an AUC of 0.72 and a TDL of 3.33, the combination of SVD and LR is found to be the top performer. Interestingly, the combination of SVD and LR outperformed more advanced ensemble modeling techniques. This emphasizes that data preparation and feature engineering mainly drive predictive performance and not the algorithm [18]. Our results are important to NPOs as they represent an alternative way to current acquisition practices. Traditionally, donor acquisition is extremely expensive because nonprofit organizations have to purchase external databases. This study shows that it is possible to build an accurate acquisition model using only Facebook data and, therefore, presents a cheaper and effective way of identifying new donors. Furthermore, our results serve as a decision support system to NPOs who want to use social media data for acquisition. Our data collection framework provides guidance as to how to gather the prospect list and the relevant social media data. Our analytical framework recommends which combination of data analytical techniques to use to construct an accurate predictive model. Based on the predictions of the resulting acquisition model, as well as the available budget, NPOs can directly determine which Facebook fans to target during acquisition campaigns.

The second objective of this study is to determine what type of Facebook features and what predictors are most important in the prediction of donation behavior. The results reveal that variables related to Facebook pages and page categories are most valuable. Variables related to Facebook groups, as well as frequency variables and gender, are found to be less important. These insights are valuable to NPOs as they provide guidance as to which type of predictors to include in acquisition models. Because collecting and preprocessing data is expensive, it could be worthwhile to implement acquisition models using only Facebook pages or Facebook page categories. With AUCs (TDLs) of 0.70 (2.92) and 0.68 (3.00) respectively, these models do not

perform significantly worse than the most complete model. Hence, our results allow NPOs to create predictive acquisition models that are not only accurate but also as efficient as possible. Finally, we provide insight into the individual characteristics that differentiate donors from non-donors. Inspection of the top dimensions show that aspects such as life stage, residence, materialism, responsible consumption, and interest in (other) NPOs are most important in predicting first-time donation behavior. Insights into these characteristics can help NPOs gain a better understanding of their online donor base and tailor appropriate marketing campaigns.

In summary, this study proves the value of social media to nonprofits. Specifically, we demonstrate how nonprofits can use Facebook (data) to identify a pool of interested individuals (i.e., Facebook fans) and predict which ones are most likely to become a donor. As such, we offer a potential solution to the current expensive acquisition practices. Furthermore, our analyses can easily be reproduced and implemented by NPOs. Our results offer guidance regarding which methodology to use, as well as which Facebook features to consider. Moreover, we hope that this study encourages nonprofit organizations to use Facebook as a tool for predictive targeted marketing rather than for simple communication.

6. Limitations and future research

This study contributes to research in donor acquisition in the nonprofit sector by creating a predictive model using social media data. Nevertheless, as this domain is relatively unexplored, several limitations and opportunities for further research arise.

A first limitation of this study is related to our data collection framework. Since we exclusively collected Facebook data of public Facebook profiles, this study may suffer from selection effects. It might be that some Facebook fans are unwilling to publicly share their data and that these Facebook fans have a different behavior than the fans in our sample. An alternative to our data collection framework would be to create a Facebook app that collects individual data with users' consent (e.g., [5]). However, due to recent scandals surrounding Facebook [59], very

few users will allow this. Also, the possibilities of such an app have been restricted considerably. Hence, organizations that want to harness the power of social media will have no other choice but to use a data collection framework similar to ours and will, therefore, be confronted to the same limitation. As such, this study is valuable to both researchers and practitioners, as it demonstrates how Facebook data can be used in this time and age.

A second limitation results from the strong class imbalance in our data set (i.e., only 1.23% of the Facebook fans are donors). To solve this problem we performed random oversampling on the training set until a 50/50 distribution was reached. However, considering the limited number of events (i.e., 96 donors) and the great number of features (i.e., 303), there will be a lot of similarity amongst the training samples, which could lead to an overestimation of the results. However, since a rigorous experimental set-up with a strict separation between training and test set on each fold was followed, we believe that our results have value and can be interpreted. Additional analyses proved that the models based on DR techniques generally do not overfit, whereas the binary models (i.e., the models not using DR) do exhibit overfitting. Furthermore, the class imbalance is inherent to social media: the number of fans and followers of NPOs is ever increasing but only very few of them actually donate, which makes resampling inevitable. Since a large benchmark study of data resampling methods is not within the scope of this study, we decided not to implement other strategies. One avenue for future research can be to study the impact of other resampling methods (e.g., SMOTE) on our results.

Third, our aim is to present a decision support system that allows NPOs to predict acquisition based on solely social media data. Hence, the acquisition models in this study are built on Facebook data only. As a topic for future research, the predictive value of social media data could be compared to that of traditional data sources (e.g., external data). In their study, Meire et al. [49] evaluate the added value of social media data over commercially purchased and website data for acquisition models in a B2B setting. A similar study can be set up in the nonprofit sector. In summary, whereas we focused on the value of different types of Facebook features, future

research could investigate the (added) value of different data sources for acquisition modeling in the nonprofit sector.

7. References

[1] Alpaydin, E. (1999). Combined 5× 2 cv F test for comparing supervised classification learning algorithms. *Neural computation*, *11*(8), 1885-1892.

[2] Althoff, T., & Leskovec, J. (2015, May). Donor retention in online crowdfunding communities: A case study of donorschoose. org. In *Proceedings of the 24th international conference on world wide web* (pp. 34-44). International World Wide Web Conferences Steering Committee.

[3] Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012, June). Personality and patterns of Facebook usage. In *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 24- 32). ACM.

[4] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, *54*(6), 627-635

[5] 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.

[6] Bekkers, R. (2006). Traditional and health-related philanthropy: The role of resources and personality. *Social psychology quarterly*, *69*(4), 349-366.

[7] Bennett, R. (2003). Factors underlying the inclination to donate to particular types of charity. *International Journal of Nonprofit and Voluntary Sector Marketing*, *8*(1), 12-29.

[8] 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.

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

[10] 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*(1-2), 501-527.

[11] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[12] Brown, S., & Taylor, K. (2015). Charitable behaviour and the big five personality traits: Evidence from UK panel data.

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

[14] Carlo, G., Okun, M. A., Knight, G. P., & de Guzman, M. R. T. (2005). The interplay of traits and motives on volunteering: Agreeableness, extraversion and prosocial value motivation. *Personality and Individual Differences*, *38*(6), 1293-1305.

[15] Chell, K., & Mortimer, G. (2014). Investigating online recognition for blood donor retention: an experiential donor value approach. *International Journal of Nonprofit and Voluntary Sector Marketing*, *19*(2), 143-163.

[16] Clark, J., & Provost, F. (2016). Matrix-Factorization-Based Dimensionality Reduction in the Predictive Modeling Process: A Design Science Perspective.

[17] Courtois, C., & Verdegem, P. (2015). Like to engage: a multi-level analysis of connective action on Facebook. In *ICA (International Communication Association), Annual conference*.

[18] Coussement, K., Lessmann, S., & Verstraeten, G. (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.

[19] Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2018). The operational value of social media information. *Production and Operations Management*, *27*(10), 1749-1769.

[20] De Caigny, A., Coussement, K., De Bock, K. W., & Lessmann, S. (2019). Incorporating textual information in customer churn prediction models based on a convolutional neural network. *International Journal of Forecasting*.

[21] D'Haen, J., Van den Poel, D., Thorleuchter, D., & Benoit, D. F. (2016). Integrating expert knowledge and multilingual web crawling data in a lead qualification system. *Decision Support Systems*, *82*, 69-78.

[22] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, *10*(7), 1895-1923.

[23] Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Springer, Berlin, Heidelberg.

[24] Enjolras, B., Steen-Johnsen, K., & Wollebæk, D. (2013). Social media and mobilization to offline demonstrations: Transcending participatory divides?. *New Media & Society*, *15*(6), 890-908.

[25] Facebook (2020). Newsroom – Key facts. http://newsroom.fb.com/Key-Facts.

[26] Farrow, H., & Yuan, Y. C. (2011). Building stronger ties with alumni through Facebook to increase volunteerism and charitable giving. *Journal of Computer-Mediated Communication*, *16*(3), 445-464.

[27] Ferguson, E. (2004). Conscientiousness, emotional stability, perceived control and the frequency, recency, rate and years of blood donor behaviour. *British Journal of Health Psychology*, *9*(3), 293-314.

[28] Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, *28*(2), 337- 407.

[29] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

[30] García, S., Fernández, A., Luengo, J., & Herrera, F. (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.

[31] Garner, J. T., & Garner, L. T. (2011). Volunteering an opinion: Organizational voice and

volunteer retention in nonprofit organizations. *Nonprofit and Voluntary Sector Quarterly*, *40*(5), 813-828.

[32] Germain, M., Glynn, S. A., Schreiber, G. B., Gélinas, S., King, M., Jones, M., ... & Tu, Y. (2007). Determinants of return behavior: a comparison of current and lapsed donors. *Transfusion*, *47*(10), 1862-1870.

[33] Godin, G., Conner, M., Sheeran, P., Bélanger-Gravel, A., & Germain, M. (2007). Determinants of repeated blood donation among new and experienced blood donors. *Transfusion*, *47*(9), 1607-1615.

[34] Harrison, J., Kim, J.Y., 2020. RSelenium: R Bindings for "Selenium WebDriver."

[35] Hsu, J. L., Liang, G. Y., & Tien, C. P. (2005). Social concerns and willingness to support charities. *Social Behavior and Personality: an international journal*, *33*(2), 189-200.

[36] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

[37] Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks. *Machine learning*, *95*(3), 357-380.

[38] Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802- 5805.

[39] Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological methods*, *21*(4), 493.

[40] Kristoffersen, L., & Singh, S. (2004). Successful application of a customer relationship management program in a nonprofit organization. *Journal of Marketing Theory and Practice*, *12*(2), 28-42.

[41] Larivière, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial

services. *Expert Systems with Applications*, *27*(2), 277-285.

[42] Lee, Y. K., & Chang, C. T. (2007). Who gives what to charity? Characteristics affecting donation behavior. *Social Behavior and Personality: an international journal*, *35*(9), 1173-1180.

[43] Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems* (pp. 556-562).

[44] Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, *43*(2), 276-286.

[45] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, *247*(1), 124-136.

[46] Lucas, E. (2017). Reinventing the rattling tin: How UK charities use Facebook in fundraising. *International Journal of Nonprofit and Voluntary Sector Marketing*, *22*(2).

[47] Mandel, J. (1982). Use of the singular value decomposition in regression analysis. *The American Statistician*, *36*(1), 15-24.

[48] Martens, D., Provost, F., Clark, J., & de Fortuny, E. J. (2016). Mining Massive Fine-Grained Behavior Data to Improve Predictive Analytics. *MIS quarterly*, *40*(4).

[49] 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*, 26-37.

[50] Meire, M., Hewett, K., Ballings, M., Kumar, V., & Van den Poel, D. (2019). The Role of Marketer-Generated Content in Customer Engagement Marketing. *Journal of Marketing*, *83*(6), 21-42.

[51] Praet, S., Van Aelst, P., & Martens, D. (2018). I like, therefore I am: predictive modeling to gain insights in political preference in a multi-party system.

[52] Sargeant, A. (1999). Charitable giving: Towards a model of donor behavior. *Journal of Marketing Management*, *15*(4), 215-238.

[53] Sargeant, A. (2001). Relationship fundraising: How to keep donors loyal. *Nonprofit Management and Leadership*, *12*(2), 177-192.

[54] Schlumpf, K. S., Glynn, S. A., Schreiber, G. B., Wright, D. J., Randolph Steele, W., Tu, Y., ... & National Heart, Lung, and Blood Institute Retrovirus Epidemiology Donor Study. (2008). Factors influencing donor return. *Transfusion*, *48*(2), 264-272.

[55] Schröder, N., Falke, A., Hruschka, H., & Reutterer, T. (2019). Analyzing the Browsing Basket: A Latent Interests-Based Segmentation Tool. *Journal of Interactive Marketing*, *47*, 181-197.

[56] Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012). Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. *Expert systems with applications*, *39*(3), 2597-2605.

[57] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288.

[58] van Dam, J. W., & van de Velden, M. (2015). Online profiling and clustering of Facebook users. *Decision Support Systems*, *70*, 60-72.

[59] Vengattil, M. (2019, September 20). Facebook suspends tens of thousands of apps in response to Cambridge Analytics row. *Reuters.* Retrieved from http://www.reuters.com

[60] Verhaert, G. A., & Van den Poel, D. (2011). Empathy as added value in predicting donation behavior. *Journal of Business Research*, *64*(12), 1288-1295.

[61] Warren, A. M., Sulaiman, A., & Jaafar, N. I. (2015). Understanding civic engagement behaviour on Facebook from a social capital theory perspective. *Behaviour & Information Technology*, *34*(2), 163-175.

[62] Zhang, Y., & Pennacchiotti, M. (2013, May). Predicting purchase behaviors from social media. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1521-1532). ACM.